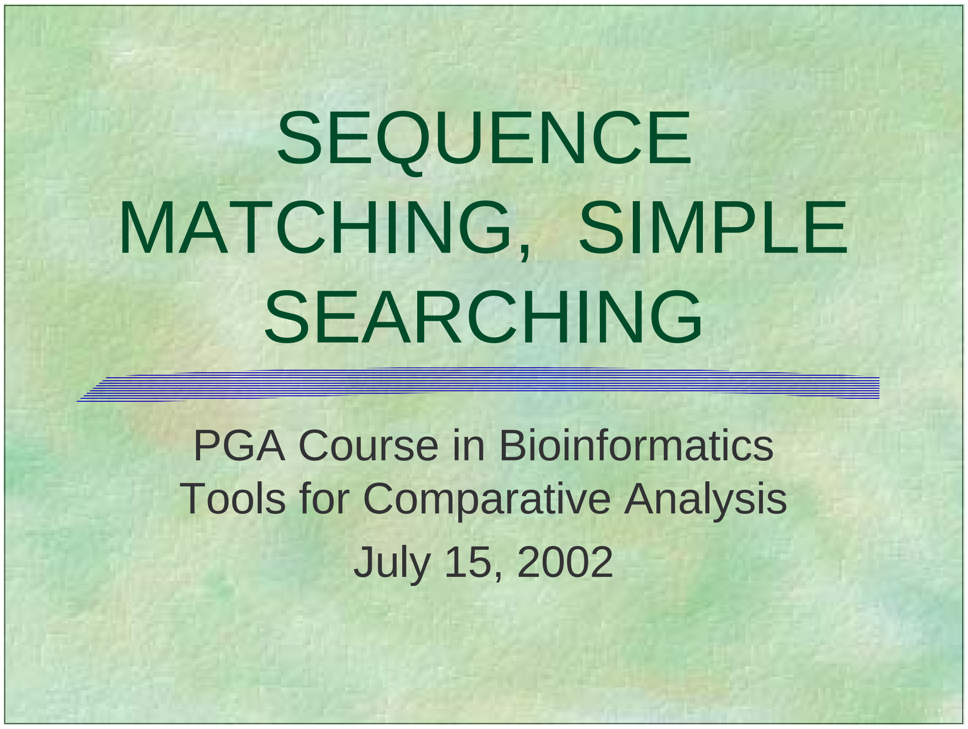
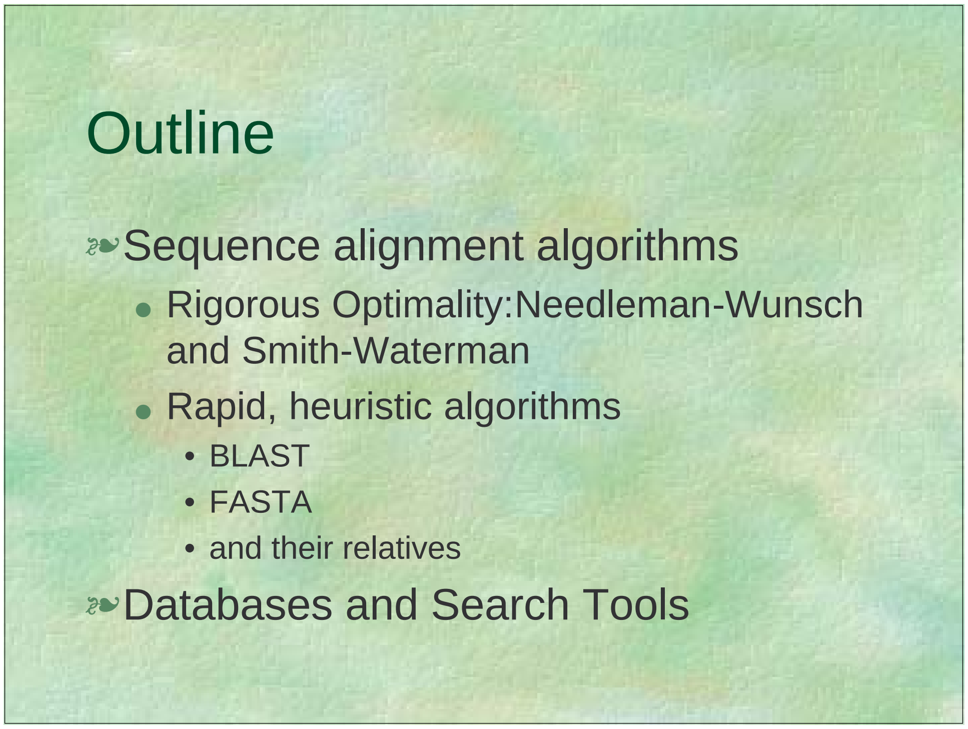


SEQUENCE MATCHING, SIMPLE SEARCHING

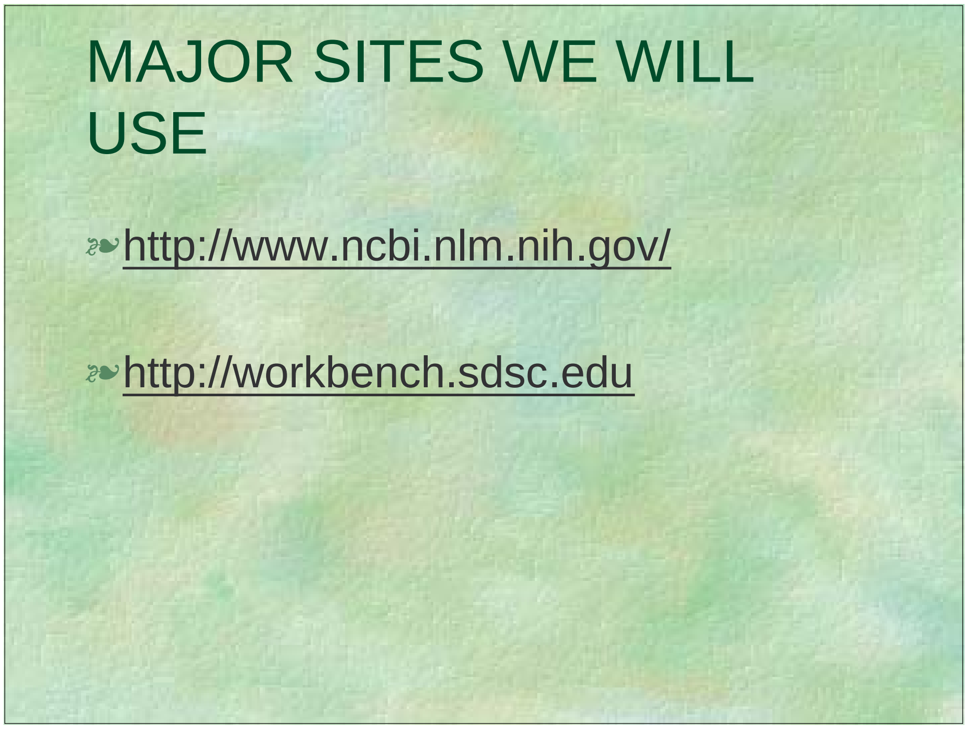
PGA Course in Bioinformatics
Tools for Comparative Analysis

July 15, 2002



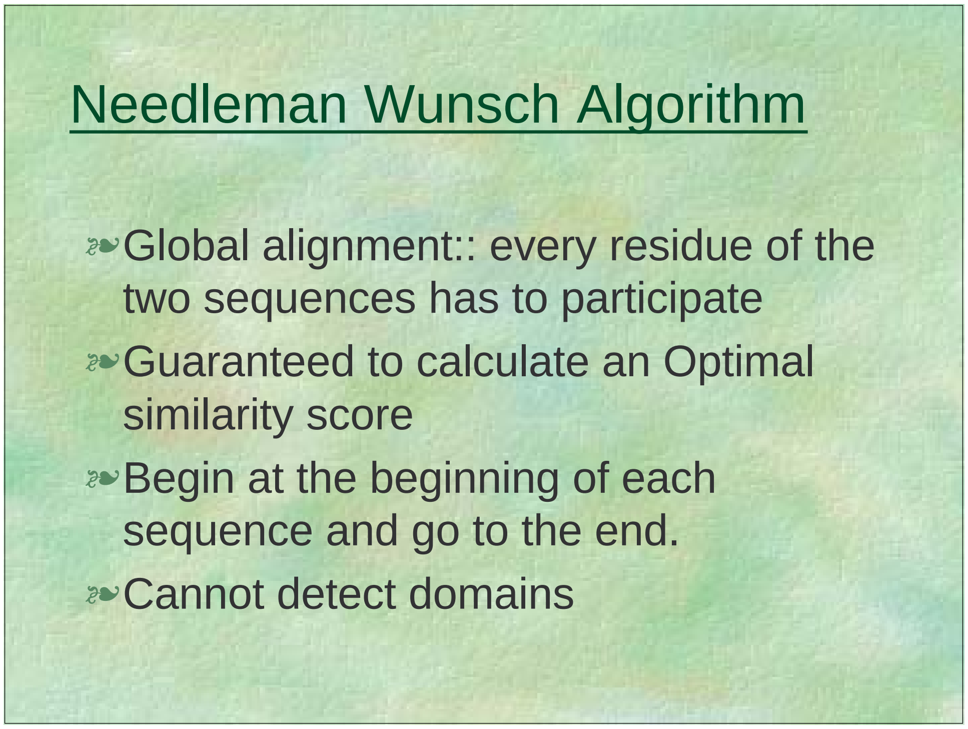
Outline

- Sequence alignment algorithms
  - Rigorous Optimality:Needleman-Wunsch and Smith-Waterman
  - Rapid, heuristic algorithms
    - BLAST
    - FASTA
    - and their relatives
- Databases and Search Tools



MAJOR SITES WE WILL USE

- http://www.ncbi.nlm.nih.gov/
- http://workbench.sdsc.edu



Needleman Wunsch Algorithm

- Global alignment:: every residue of the two sequences has to participate
- Guaranteed to calculate an Optimal similarity score
- Begin at the beginning of each sequence and go to the end.
- Cannot detect domains

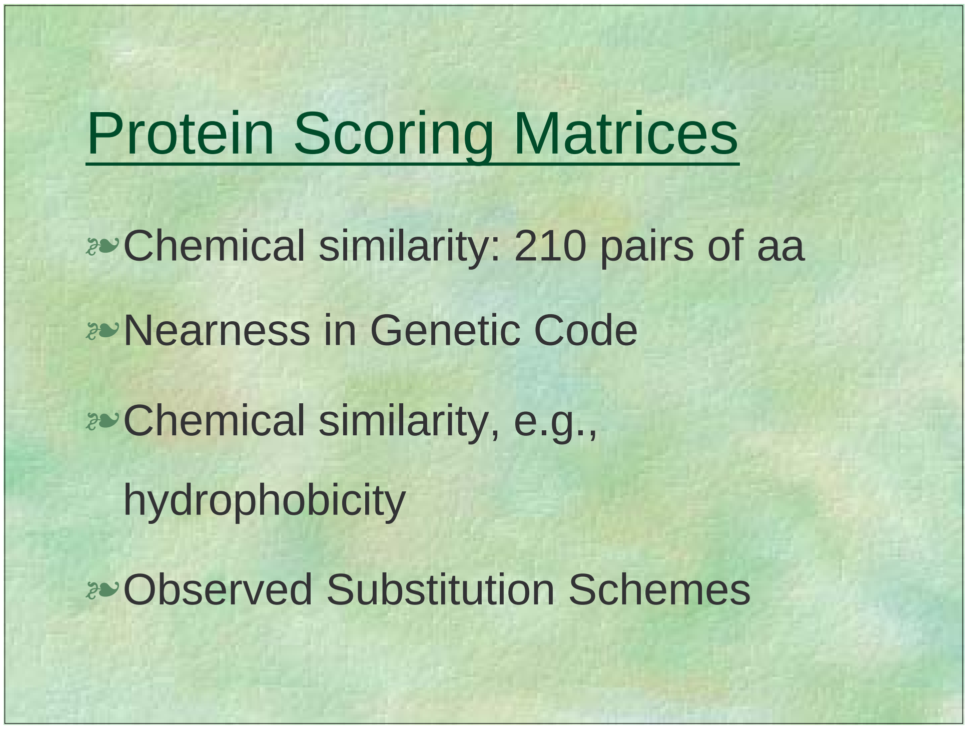
Smith-Waterman Algorithm

- Optimal Local Alignment
- Guaranteed to find all signficant matches to a given query
- Takes the query sequence versus every sequence in the database
- Can be used with arbitrary scoring systems
- **COMPUTATIONALLY EXPENSIVE!!!**

Scoring Matrices

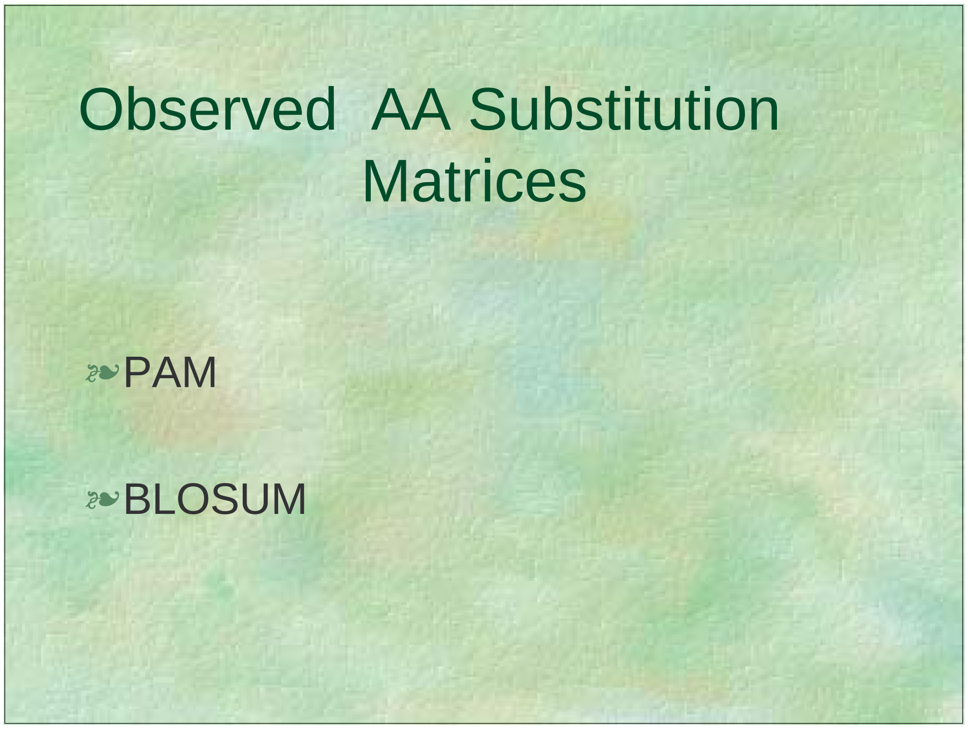
- Relatively simple for DNA-gap penalties or mismatches-can be made to look at Pu/Py

- Protein matches look also at similarity (leu/ileu)



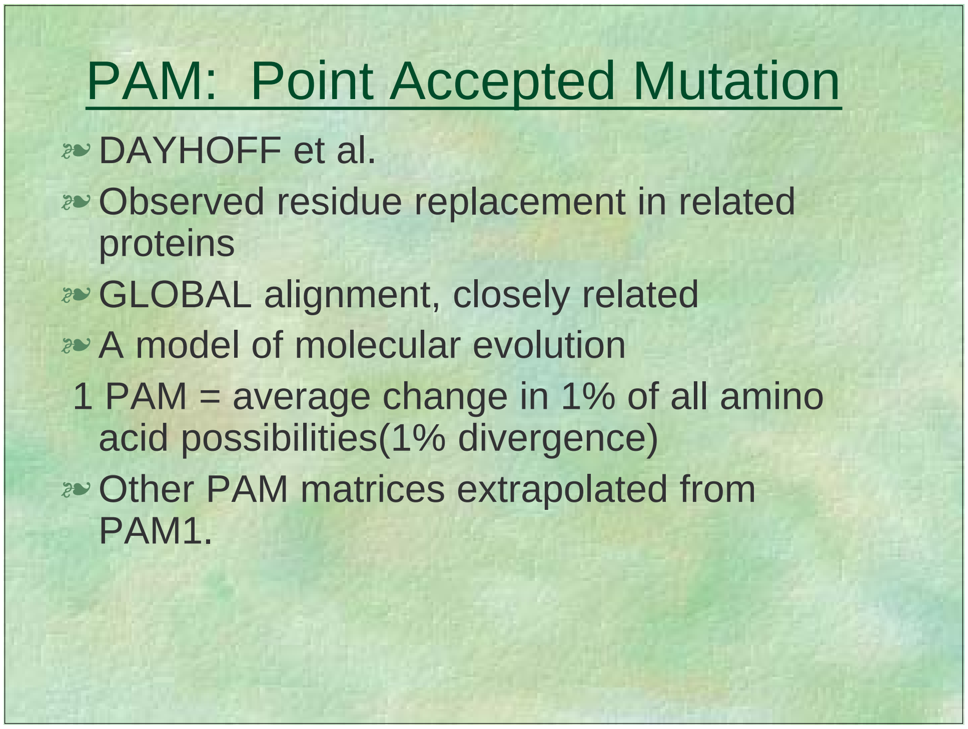
Protein Scoring Matrices

- Chemical similarity: 210 pairs of aa
- Nearness in Genetic Code
- Chemical similarity, e.g., hydrophobicity
- Observed Substitution Schemes



Observed AA Substitution Matrices

- PAM
- BLOSUM

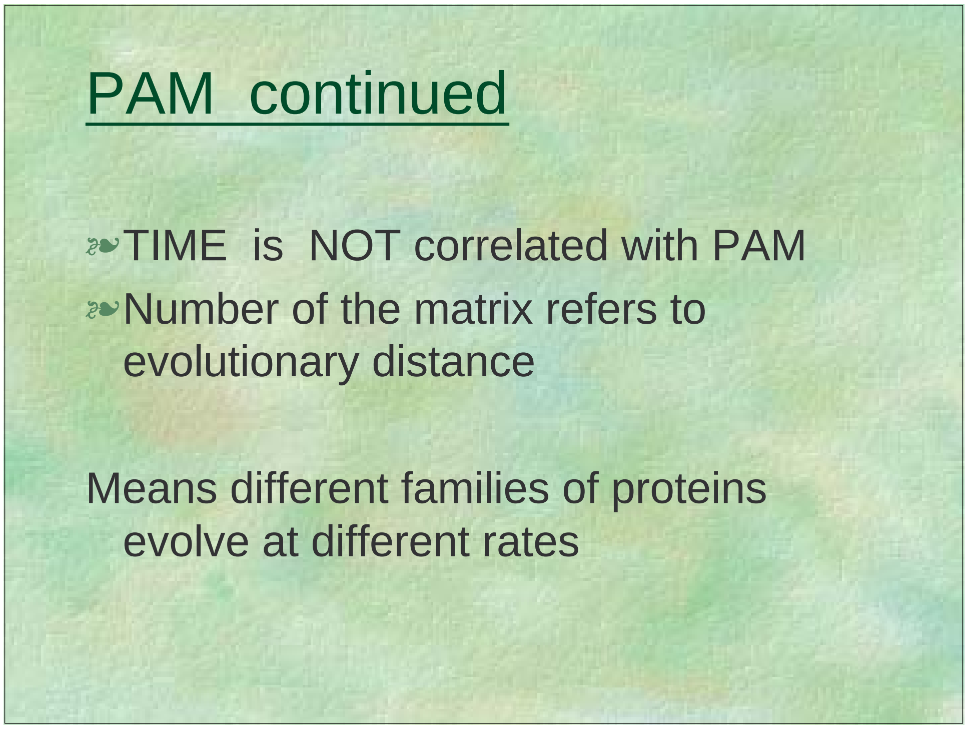


PAM: Point Accepted Mutation

- DAYHOFF et al.
- Observed residue replacement in related proteins
- GLOBAL alignment, closely related
- A model of molecular evolution

1 PAM = average change in 1% of all amino acid possibilities(1% divergence)

- Other PAM matrices extrapolated from PAM1.



PAM continued

- TIME is NOT correlated with PAM
- Number of the matrix refers to evolutionary distance

Means different families of proteins evolve at different rates

PAM250

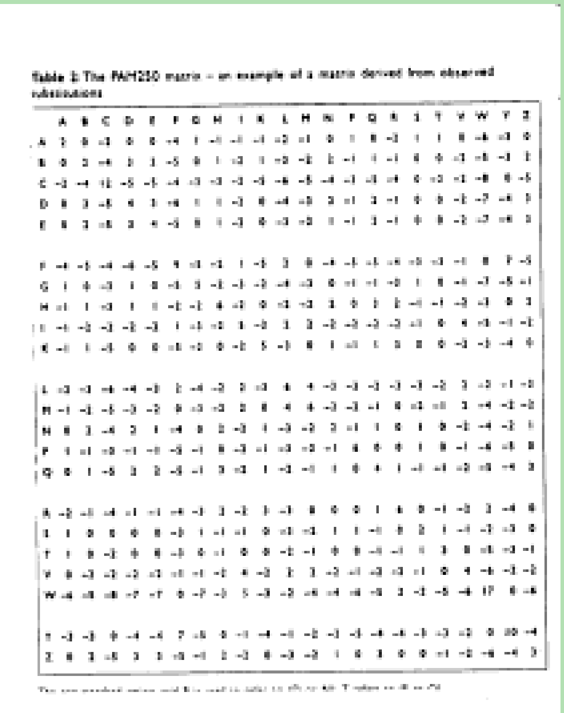
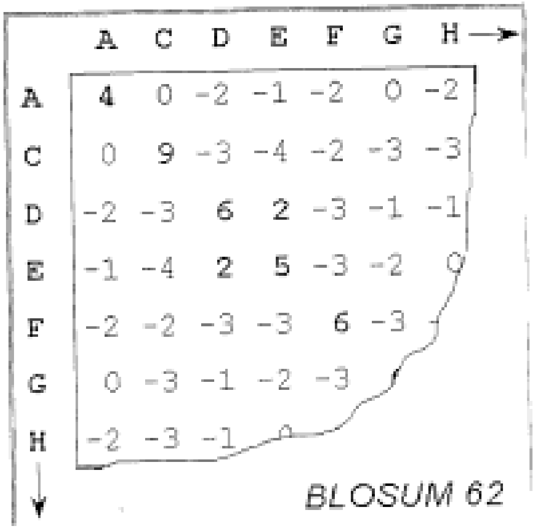


Table 1: The PAM250 matrix – an example of a matrix derived from observed substitutions

	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
A	2	0	-2	0	0	-4	1	-1	-1	-1	-2	-1	0	1	0	-2	1	1	0	-6	-3	0
B	0	2	-4	3	3	-5	0	1	-2	1	-3	-2	2	-1	1	-1	0	0	-2	-5	-3	2
C	-2	-4	12	-5	-5	-4	-3	-3	-2	-5	-6	-5	-4	-3	-5	-4	0	-2	-2	-8	0	-5
D	0	3	-5	4	3	-6	1	1	-2	0	-4	-3	2	-1	2	-1	0	0	-2	-7	-4	3
E	0	3	-5	3	4	-5	0	1	-2	0	-3	-2	1	-1	2	-1	0	0	-2	-7	-4	3
F	-4	-5	-4	-6	-5	9	-5	-2	1	-5	2	0	-4	-5	-5	-4	-3	-3	-1	0	7	-5
G	1	0	-3	1	0	-5	5	-2	-3	-2	-4	-3	0	-1	-1	-3	1	0	-1	-7	-5	-1
H	-1	1	-3	1	1	-2	-2	6	-2	0	-2	-2	2	0	3	2	-1	-1	-2	-3	0	2
I	-1	-2	-2	-2	-2	1	-3	-2	5	-2	2	2	-2	-2	-2	-2	-1	0	4	-5	-1	-2
K	-1	1	-5	0	0	-5	-2	0	-2	5	-3	0	1	-1	1	3	0	0	-2	-3	-4	0
L	-2	-3	-6	-4	-3	2	-4	-2	2	-3	6	4	-3	-3	-2	-3	-3	-2	2	-2	-1	-3
M	-1	-2	-5	-3	-2	0	-3	-2	2	0	4	6	-2	-2	-1	0	-2	-1	2	-4	-2	-2
N	0	2	-4	2	1	-4	0	2	-2	1	-3	-2	2	-1	1	0	1	0	-2	-4	-2	1
P	1	-1	-3	-1	-1	-5	-1	0	-2	-1	-3	-2	-1	6	0	0	1	0	-1	-6	-5	0
Q	0	1	-5	2	2	-5	-1	3	-2	1	-2	-1	1	0	4	1	-1	-1	-2	-5	-4	3
R	-2	-1	-4	-1	-1	-4	-3	2	-2	3	-3	0	0	0	1	6	0	-1	-2	2	-4	0
S	1	0	0	0	0	-3	1	-1	-1	0	-3	-2	1	1	-1	0	2	1	-1	-2	-3	0
T	1	0	-2	0	0	-3	0	-1	0	0	-2	-1	0	0	-1	-1	1	3	0	-5	-3	-1
V	0	-2	-2	-2	-2	-1	-1	-2	4	-2	2	2	-2	-1	-2	-2	-1	0	4	-6	-2	-2
W	-6	-5	-8	-7	-7	0	-7	-3	-5	-3	-2	-4	-4	-6	-5	2	-2	-5	-6	17	0	-6
Y	-3	-3	0	-4	-4	7	-5	0	-1	-4	-1	-2	-2	-5	-4	-4	-3	-3	-2	0	10	-4
Z	0	2	-5	3	3	-5	-1	2	-2	0	-3	-2	1	0	3	0	0	-1	-2	-6	-4	3

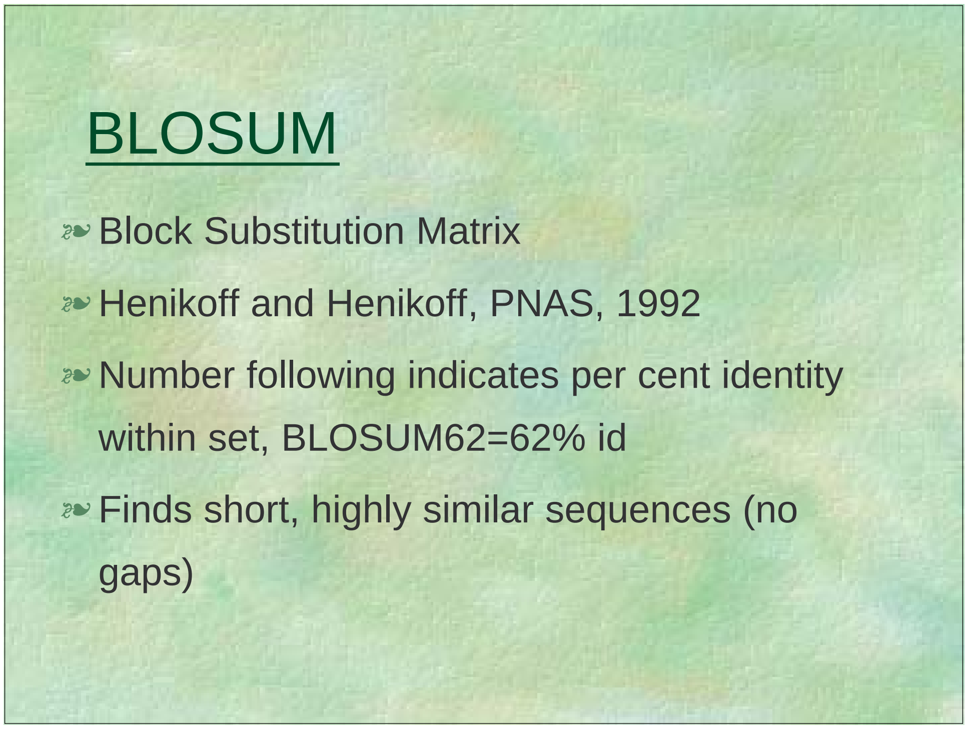
The [illegible] B is used to refer to [illegible] Z refers to [illegible]

Blosum62



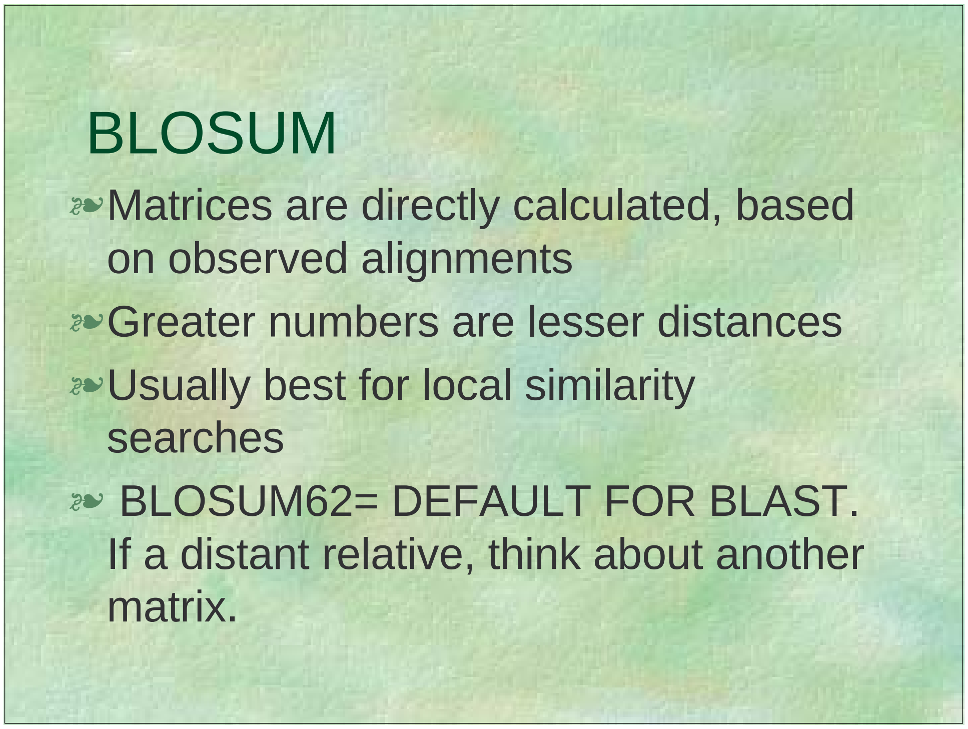
	A	C	D	E	F	G	H →
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	
G	0	-3	-1	-2	-3		
H	-2	-3	-1	0			
↓							

BLOSUM 62



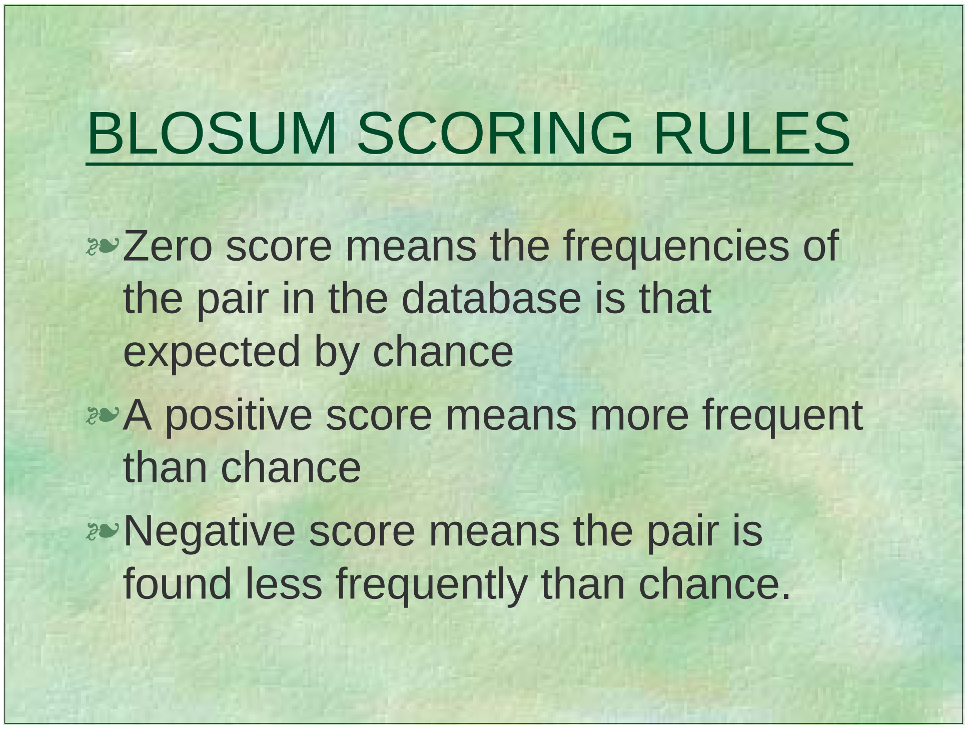
BLOSUM

- Block Substitution Matrix
- Henikoff and Henikoff, PNAS, 1992
- Number following indicates per cent identity within set, BLOSUM62=62% id
- Finds short, highly similar sequences (no gaps)



BLOSUM

- Matrices are directly calculated, based on observed alignments
- Greater numbers are lesser distances
- Usually best for local similarity searches
- BLOSUM62= DEFAULT FOR BLAST. If a distant relative, think about another matrix.

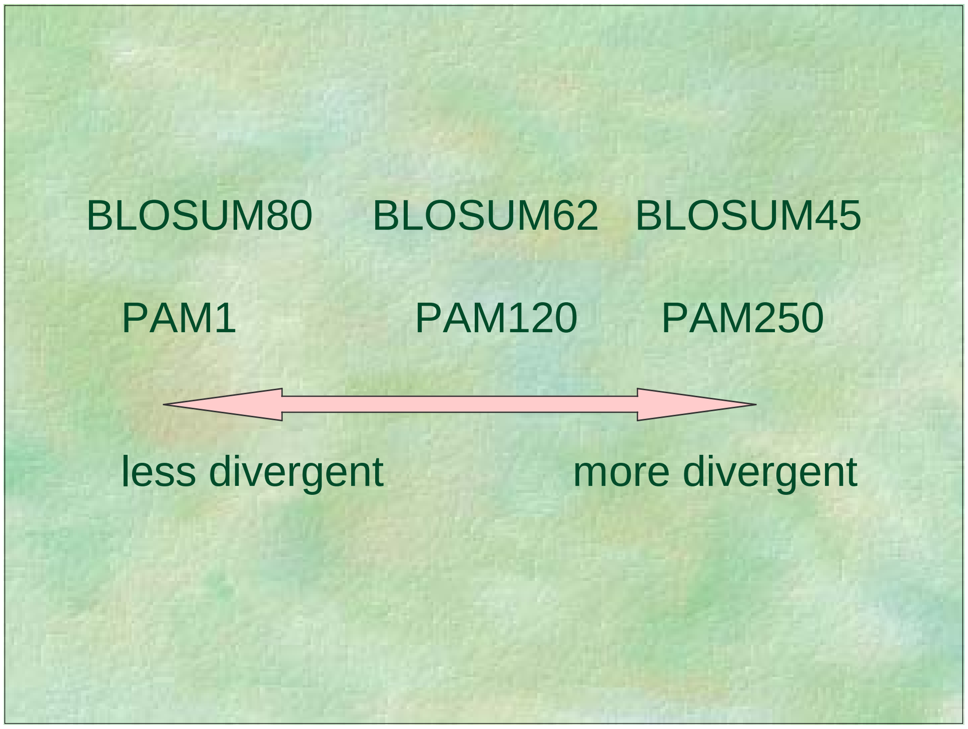


BLOSUM SCORING RULES

- Zero score means the frequencies of the pair in the database is that expected by chance
- A positive score means more frequent than chance
- Negative score means the pair is found less frequently than chance.

BLAST-Basic Local Alignment Sequence Tool

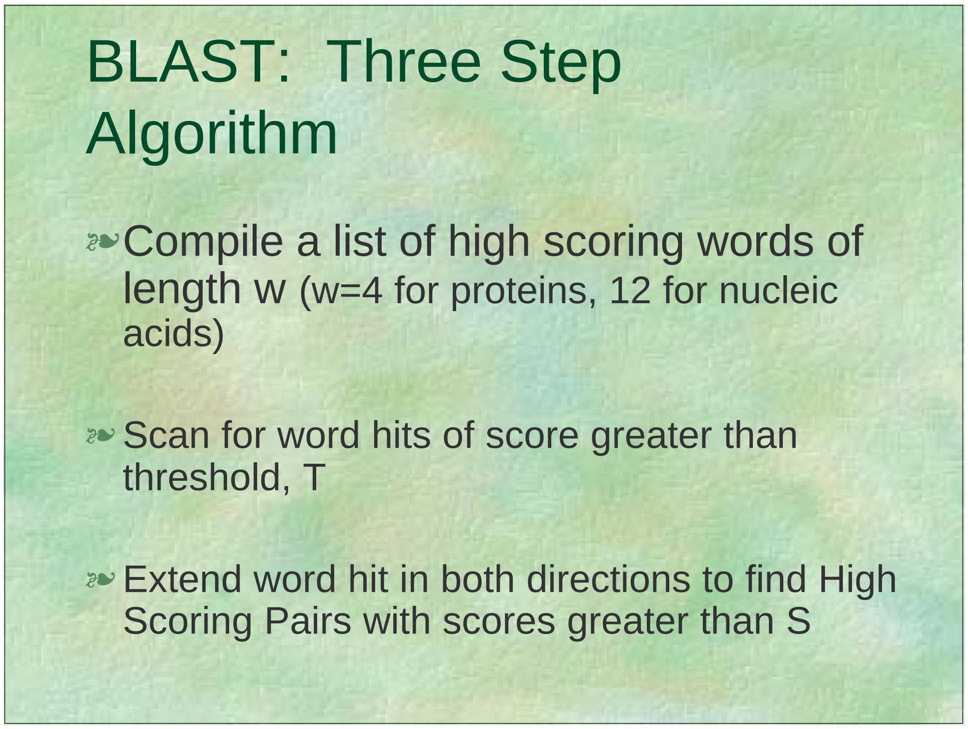
- Objective: find all local regions of similarity distinguishable from random
- Only local alignments permitted,
- Gaps permitted in version 2
- Statistically sound (Karlin and Altschul), but no guantee of optimality



BLOSUM80 BLOSUM62 BLOSUM45

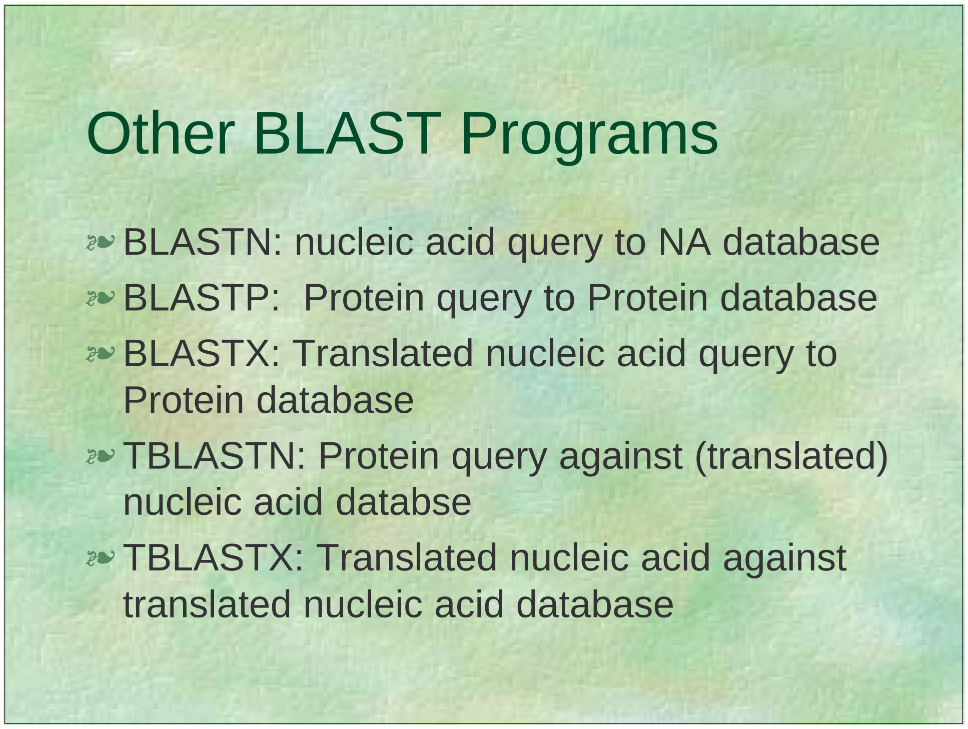
PAM1 PAM120 PAM250

less divergent more divergent



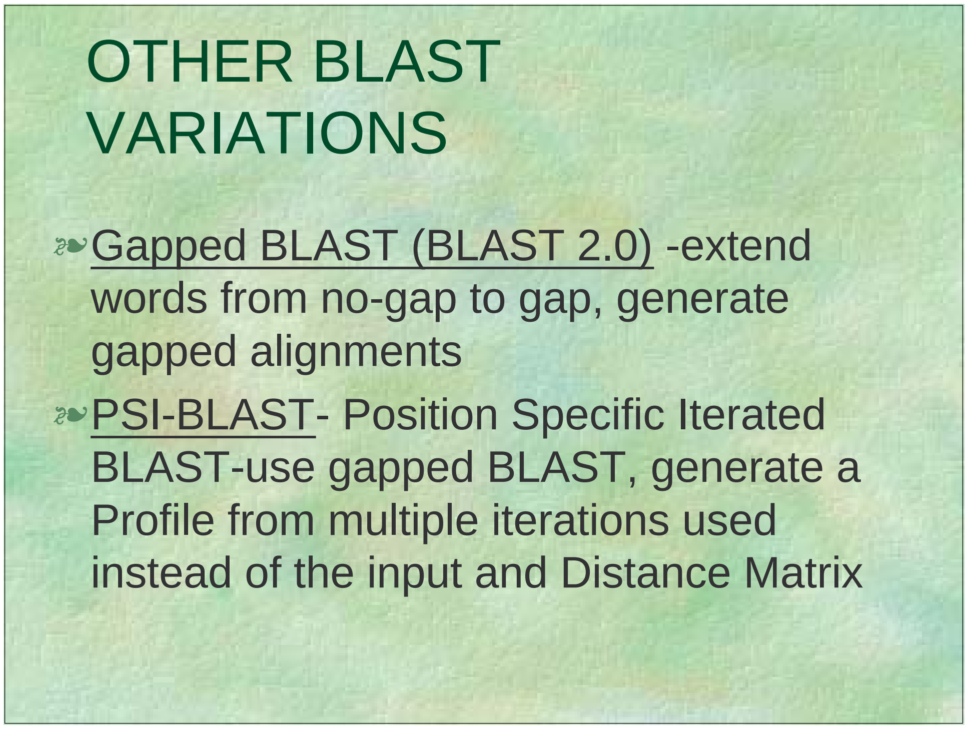
BLAST: Three Step Algorithm

- Compile a list of high scoring words of length w (w=4 for proteins, 12 for nucleic acids)
- Scan for word hits of score greater than threshold, T
- Extend word hit in both directions to find High Scoring Pairs with scores greater than S



Other BLAST Programs

- BLASTN: nucleic acid query to NA database
- BLASTP: Protein query to Protein database
- BLASTX: Translated nucleic acid query to Protein database
- TBLASTN: Protein query against (translated) nucleic acid databse
- TBLASTX: Translated nucleic acid against translated nucleic acid database



OTHER BLAST VARIATIONS

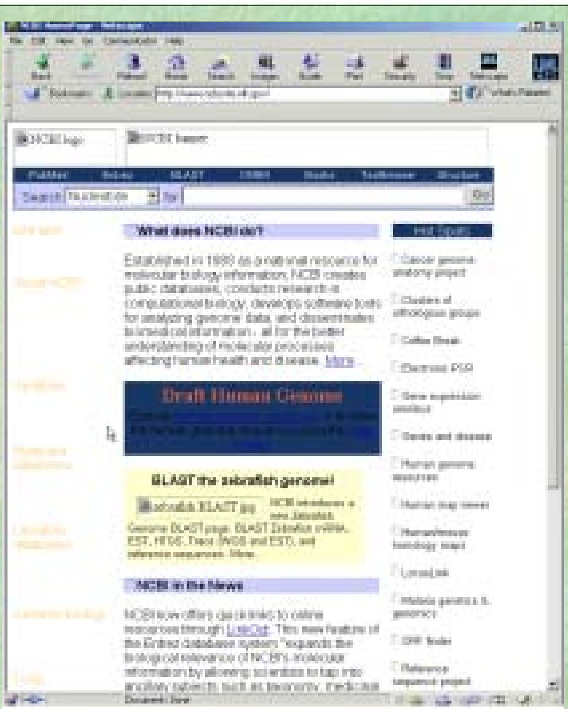
- Gapped BLAST (BLAST 2.0) -extend words from no-gap to gap, generate gapped alignments
- PSI-BLAST- Position Specific Iterated BLAST-use gapped BLAST, generate a Profile from multiple iterations used instead of the input and Distance Matrix

Limitations to BLAST

- Needs islands of strong homology
- Limits on the combination of scoring and penalty values
- The variants (blastx, tblastn, tblastx) use 6-frame translation-miss sequences with frameshifts)
- Finds and reports ONLY local alignments

A WALK THROUGH BLAST

NCBI home


NCBI HomePage - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Images Guide Print Security Shop Stop
Bookmarks Location: http://www.ncbi.nlm.nih.gov/ What's Related
NCBI logo
NCBI banner
PubMed Entrez BLAST OMIM Books TaxBrowser Structure
Search Nucleotide for Go
Hot Spots
Document: Done

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

Draft Human Genome

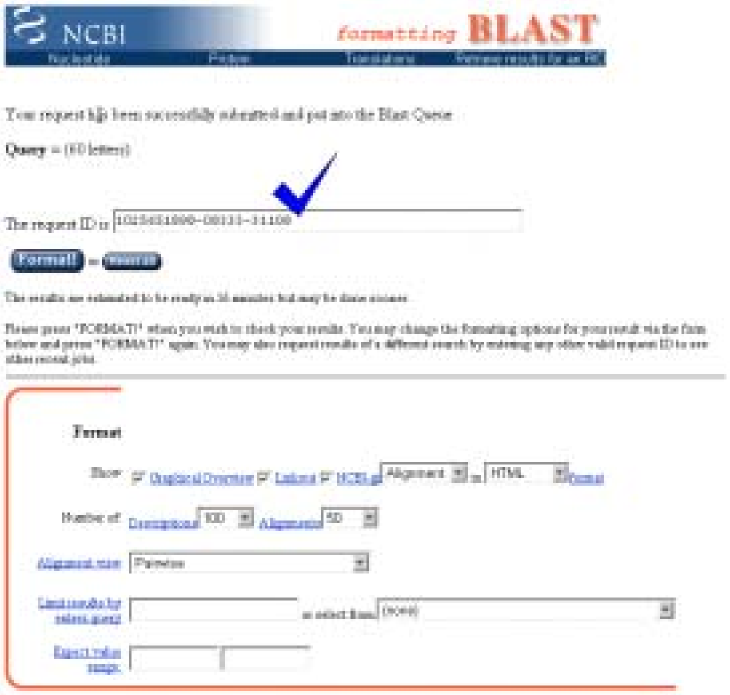
BLAST the zebrafish genome!

zebrafish BLAST.jpg NCBI introduces a new Zebrafish Genome BLAST page. BLAST Zebrafish mRNA, EST, HTGS, Trace (WGS and EST), and reference sequences. More

NCBI in the News

NCBI now offers quick links to online resources through LinkOut. This new feature of the Entrez database system "expands the biological relevance of NCBI's molecular information by allowing scientists to tap into ancillary subjects such as taxonomy, medical

- Cancer genome anatomy project
- Clusters of orthologous groups
- Coffee Break
- Electronic PCR
- Gene expression omnibus
- Genes and disease
- Human genome resources
- Human map viewer
- Human/mouse homology maps
- LocusLink
- Malaria genetics & genomics
- ORF finder
- Reference sequence project



Your request has been successfully submitted and put into the Blast Queue.

Query = (60 letters)

The request ID is 1025651890-08133-31108

The results are estimated to be ready in 16 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

Format

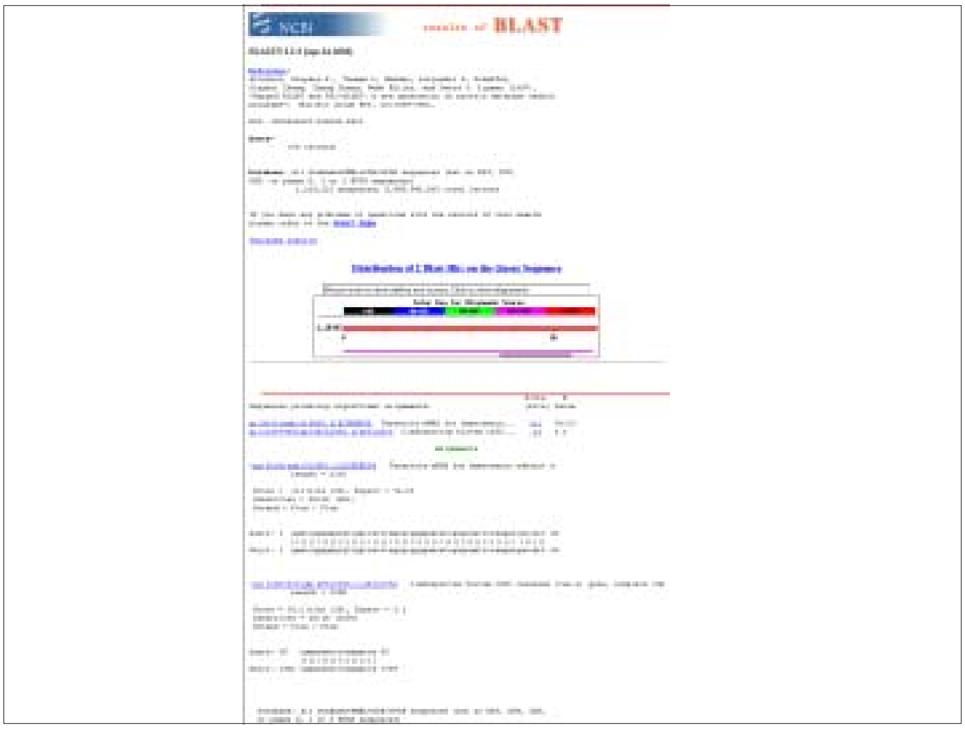
Show ☑ Graphical Overview ☑ Linkout ☑ NCBI-gi Alignment as HTML Format

Number of: Descriptions 100 Alignments 50

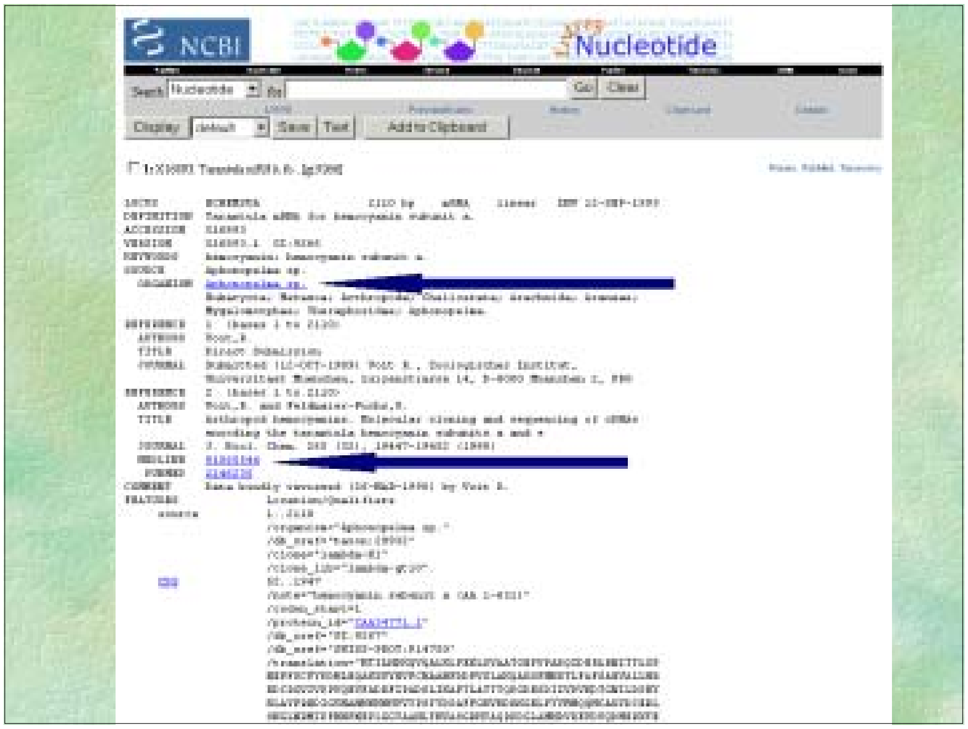
Alignment view Pairwise

Limit results by entrez query or select from: (none)

Expect value range

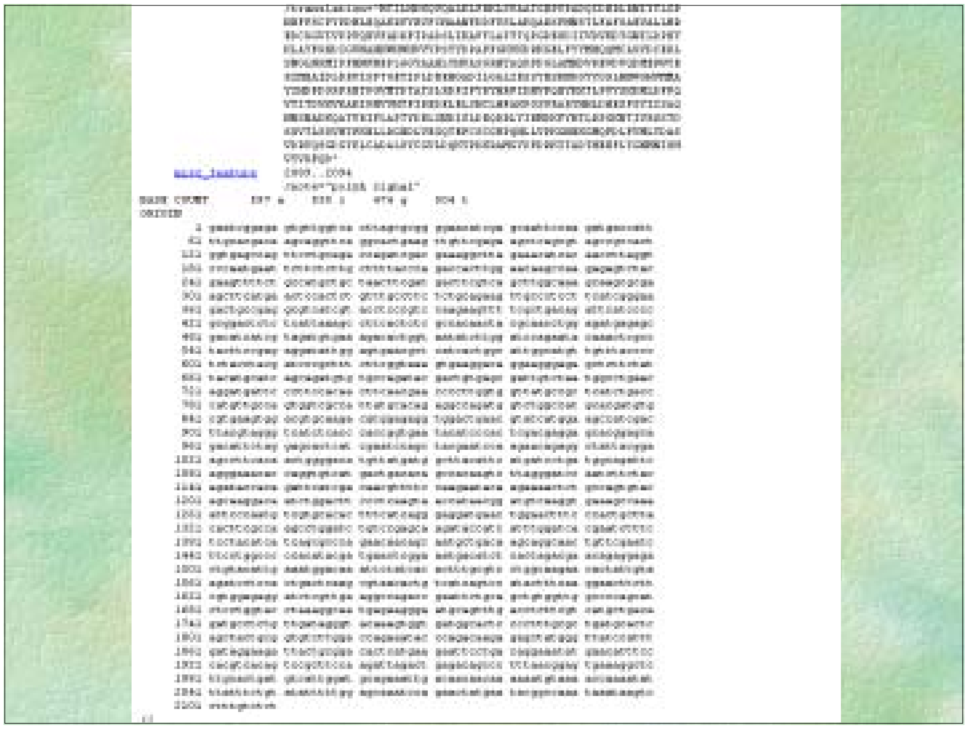

NCBI
results of BLAST

NCBI
results of BLAST

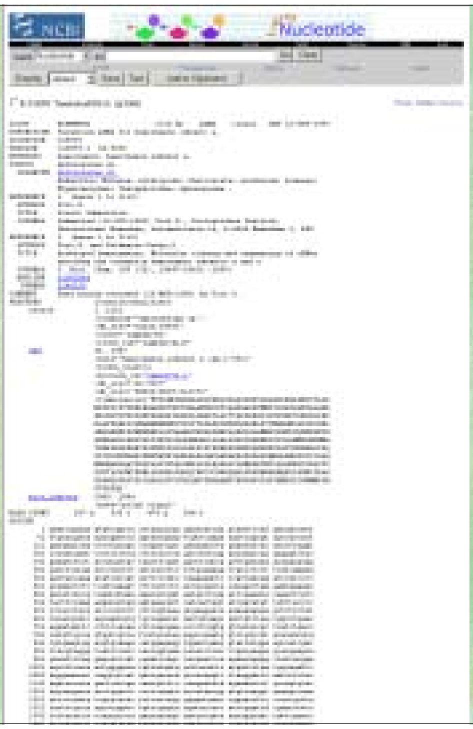


```
□ 1: X16893 Tarantula mRNA fo...[gi:9266]

LOCUS       [illegible]              2120 bp    mRNA    linear   INV [illegible]
DEFINITION  Tarantula mRNA for hemocyanin subunit a.
ACCESSION   [illegible]
VERSION     [illegible]  GI:[illegible]
KEYWORDS    hemocyanin; hemocyanin subunit a.
SOURCE      Aphonopelma sp.
  ORGANISM  Aphonopelma sp.
            Eukaryota; Metazoa; Arthropoda; Chelicerata; Arachnida; Araneae;
            Mygalomorphae; Theraphosidae; Aphonopelma.
REFERENCE   1  (bases 1 to 2120)
  AUTHORS   Voit,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (11-OCT-1989) Voit R., Zoologisches Institut,
            Universitaet Muenchen, Luisenstrasse 14, D-8000 Muenchen 2, FRG
REFERENCE   2  (bases 1 to 2120)
  AUTHORS   Voit,R. and Feldmaier-Fuchs,G.
  TITLE     Arthropod hemocyanins. Molecular cloning and sequencing of cDNAs
            encoding the tarantula hemocyanin subunits a and e
  JOURNAL   J. Biol. Chem. 265 (32), 19447-19452 (1990)
  MEDLINE   [illegible]
   PUBMED   [illegible]
COMMENT     Data kindly reviewed ([illegible]-1990) by Voit R.
FEATURES             Location/Qualifiers
     source          1..2120
                     /organism="Aphonopelma sp."
                     /db_xref="taxon:[illegible]"
                     /clone="lambda-[illegible]"
                     /clone_lib="lambda-gt10"
     CDS             [illegible]..1947
                     /note="hemocyanin subunit a (AA 1-[illegible])"
                     /codon_start=1
                     /protein_id="[illegible]"
                     /db_xref="GI:[illegible]"
                     /db_xref="SWISS-PROT:[illegible]"
                     /translation="[illegible]
                     [illegible]
```



```
                     /translation="[illegible]
                     [illegible]
                     [illegible]"
     misc_feature    [illegible]..2004
                     /note="polyA signal"
BASE COUNT      [illegible] a    [illegible] c    474 g    [illegible] t
ORIGIN
        1 [illegible]
       61 [illegible]
      [illegible]
     2101 [illegible]
//
```



BLAST RULES OF THUMB

- For short amino acid sequences (20-40), 50% identity happens by chance
- If A and B are homologous, and B and C are homologous, then A and C are, even if you can’t see it.
- You can get similarity in the absence of homology for low complexity, transmembrane and coiled-coil regions. These have to be eliminated by you.

BLAST Significance

- If you change scoring systems, you can still compare search results if you normalize the score.

  S’=(lambdaS-lnK)/ln2. Lambda and K are associated with the scoring system.

  S’, with a given E, is significant if it is greater than log N/E, N the size of the search space.

FASTA: FAST Alignment

- http://alpha10.bioch.virginia.edu/fasta/
- http://www2.ebi.ac.uk/fasta3
- http://workbench.sdsc.edu

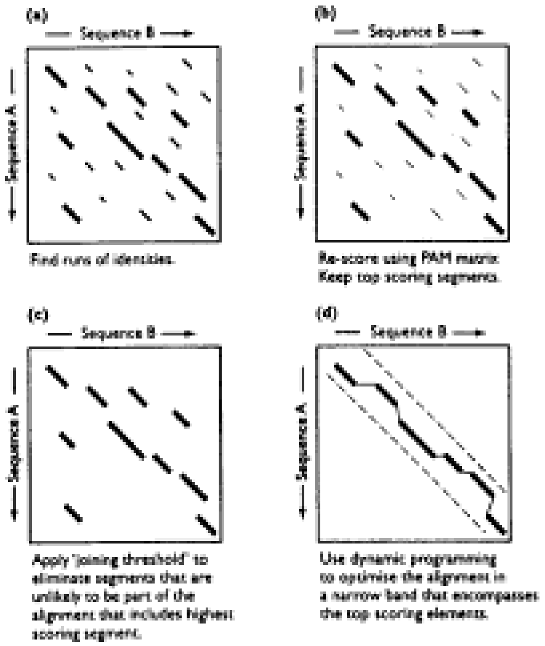
- Rapid Global alignment

- Not a strong mathematical basis

FASTA: WHY USE IT?

- Allow alignments to shift frames

FASTA Algorithm

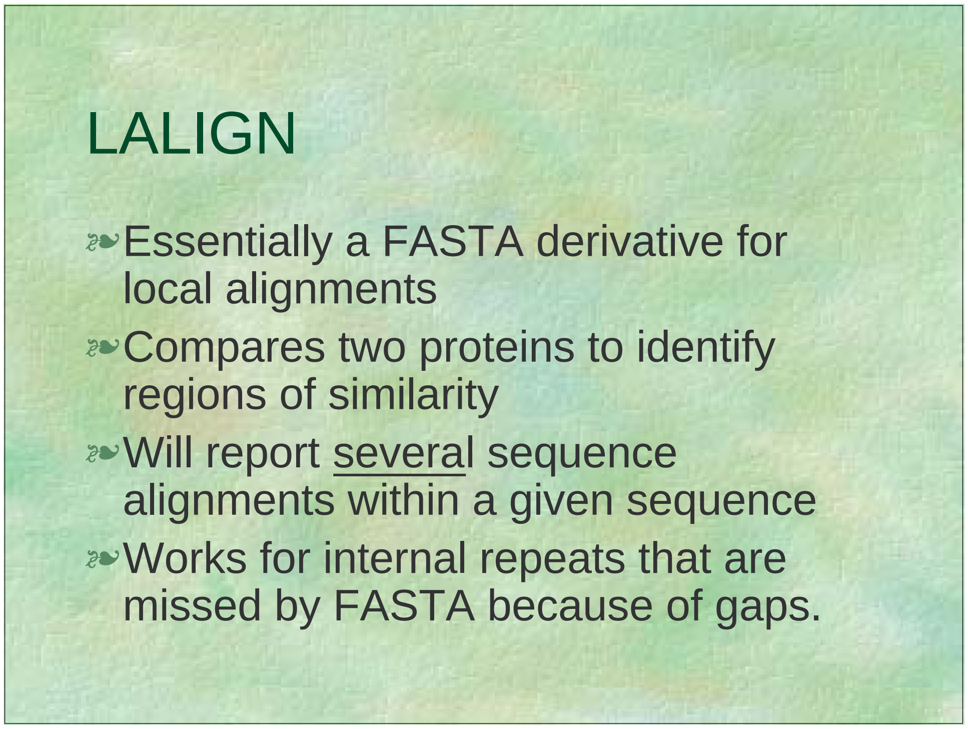


Find runs of identities.

Re-score using PAM matrix. Keep top scoring segments.

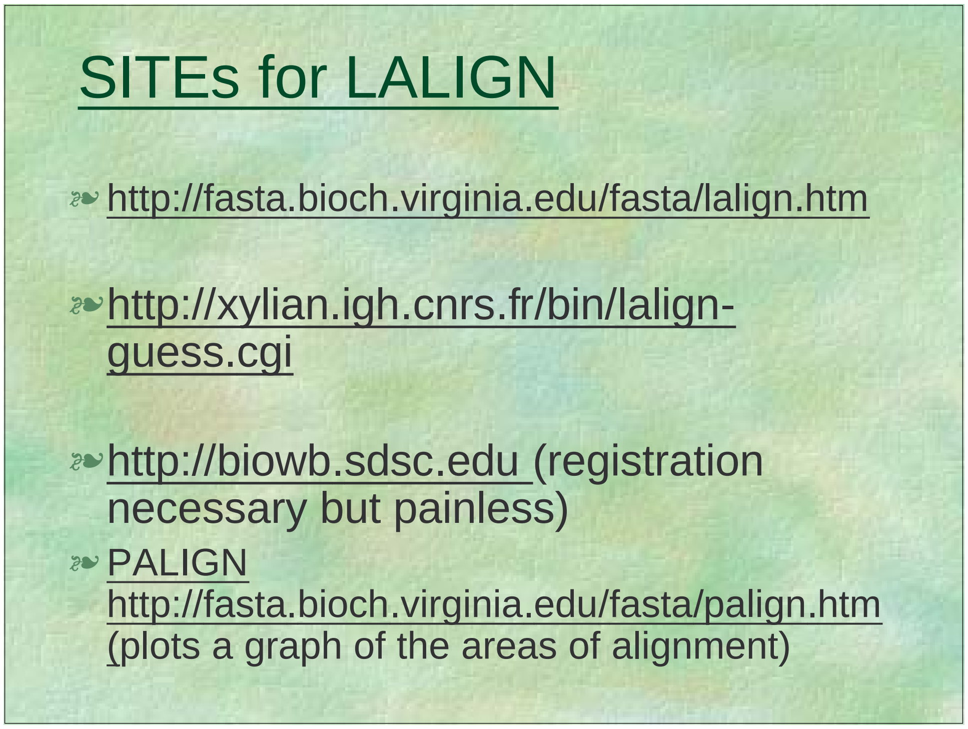
Apply 'joining threshold' to eliminate segments that are unlikely to be part of the alignment that includes highest scoring segment.

Use dynamic programming to optimise the alignment in a narrow band that encompasses the top scoring elements.



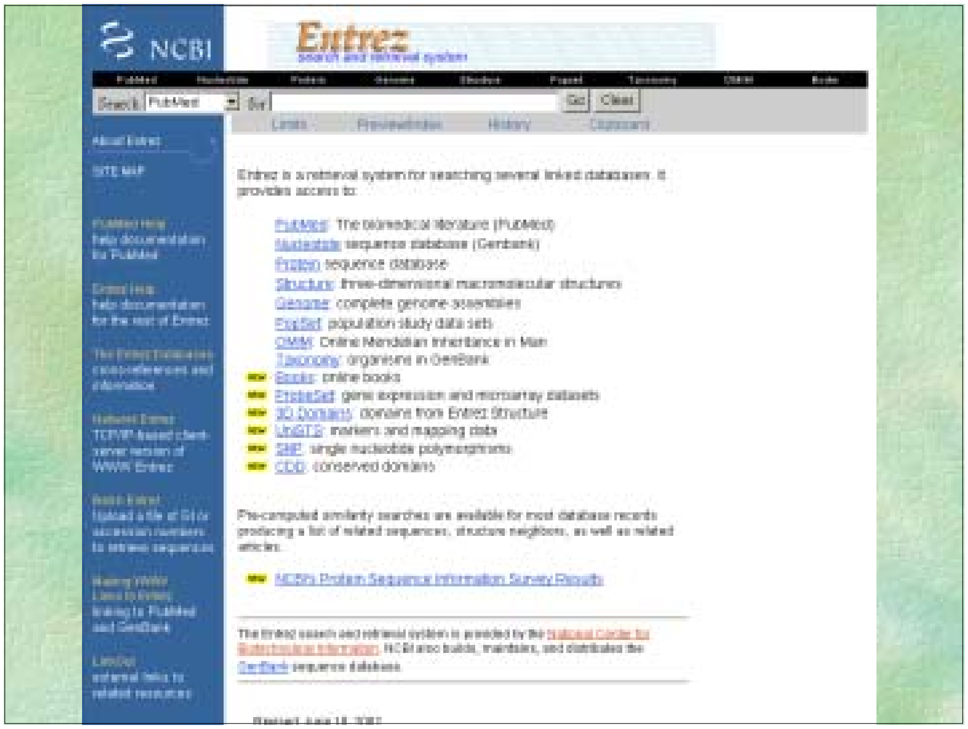
LALIGN

- Essentially a FASTA derivative for local alignments
- Compares two proteins to identify regions of similarity
- Will report several sequence alignments within a given sequence
- Works for internal repeats that are missed by FASTA because of gaps.



SITEs for LALIGN

- http://fasta.bioch.virginia.edu/fasta/lalign.htm
- http://xylian.igh.cnrs.fr/bin/lalign-guess.cgi
- http://biowb.sdsc.edu (registration necessary but painless)
- PALIGN http://fasta.bioch.virginia.edu/fasta/palign.htm (plots a graph of the areas of alignment)


NCBI
Entrez
search and retrieval system
PubMed
Nucleotide
Protein
Genome
Structure
PopSet
Taxonomy
OMIM
Books
Search
PubMed
for
Go
Clear
Limits
Preview/Index
History
Clipboard
About Entrez
SITE MAP
PubMed Help
help documentation for PubMed
Entrez Help
help documentation for the rest of Entrez
The Entrez Databases
cross-references and information
Network Entrez
TCP/IP-based client-server version of WWW Entrez
Batch Entrez
Upload a file of GI or accession numbers to retrieve sequences
Making Web Links to Entrez
linking to PubMed and GenBank
LinkOut
external links to related resources

Entrez is a retrieval system for searching several linked databases. It provides access to:

- PubMed: The biomedical literature (PubMed)
- Nucleotide: sequence database (Genbank)
- Protein: sequence database
- Structure: three-dimensional macromolecular structures
- Genome: complete genome assemblies
- PopSet: population study data sets
- OMIM: Online Mendelian Inheritance in Man
- Taxonomy: organisms in GenBank
- new Books: online books
- new ProbeSet: gene expression and microarray datasets
- new 3D Domains: domains from Entrez Structure
- new UniSTS: markers and mapping data
- new SNP: single nucleotide polymorphisms
- new CDD: conserved domains

Pre-computed similarity searches are available for most database records producing a list of related sequences, structure neighbors, as well as related articles.

new NCBI's Protein Sequence Information Survey Results

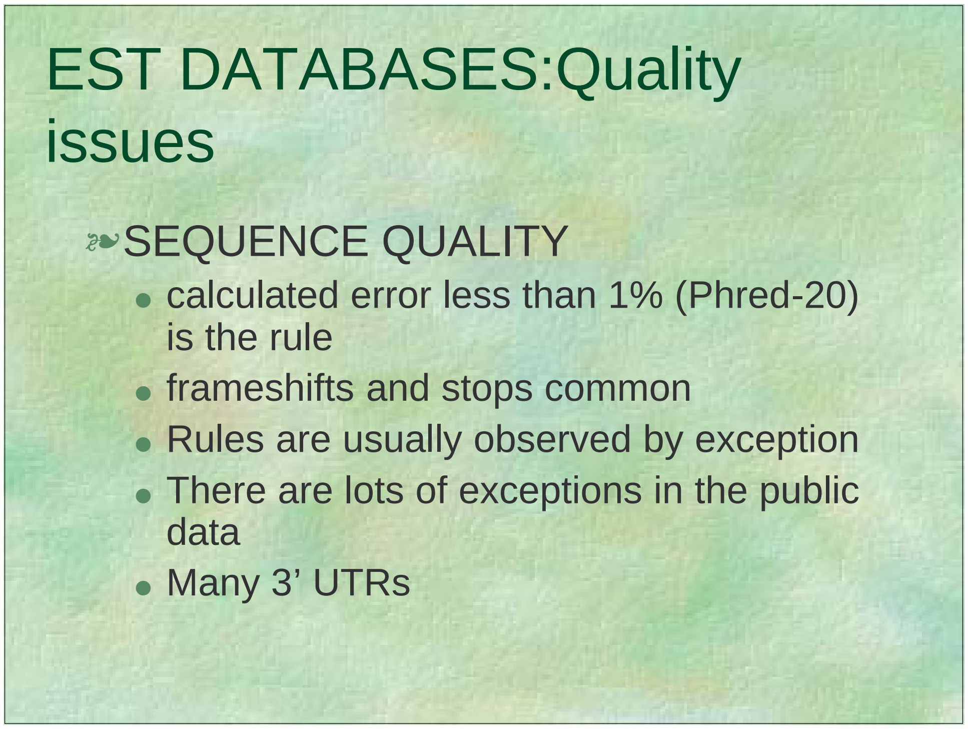
The Entrez search and retrieval system is provided by the National Center for Biotechnology Information. NCBI also builds, maintains, and distributes the GenBank sequence database.

Revised June 18, 2002

ENTREZ: Linked Databases

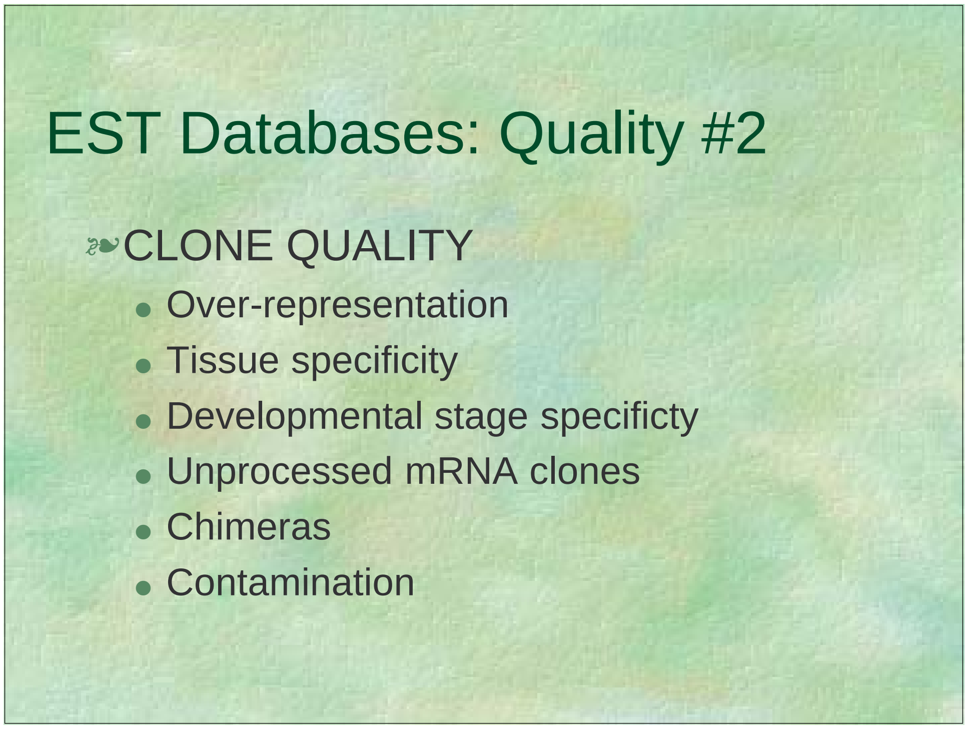
http://www.ncbi.nlm.nih.gov/Entrez/

- Concept of Neighbor-usually BLAST relationship
- Precomputed=Fast
- Related sequence, structure neighbors, related articles



EST DATABASES:Quality issues

- SEQUENCE QUALITY
  - calculated error less than 1% (Phred-20) is the rule
  - frameshifts and stops common
  - Rules are usually observed by exception
  - There are lots of exceptions in the public data
  - Many 3’ UTRs

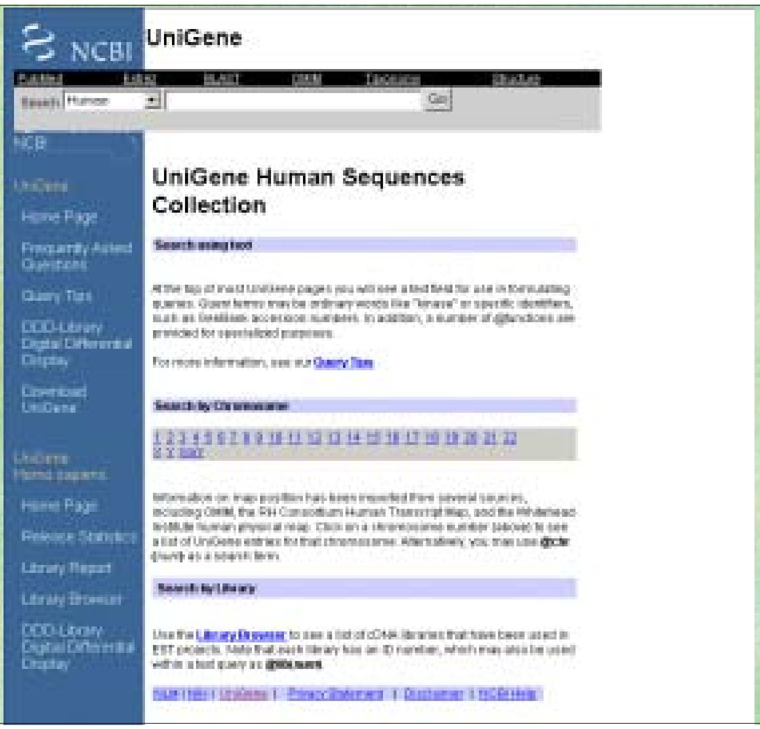


EST Databases: Quality #2

- CLONE QUALITY
  - Over-representation
  - Tissue specificity
  - Developmental stage specificty
  - Unprocessed mRNA clones
  - Chimeras
  - Contamination

EST Cluster Databases

- STACK-at SANBI http://sanbi.ac.za
- TIGR-animals, plants, other http://www.tigr.org/tdb/tgi.shtml
- Unigene-NCBI
  - Human, mouse, rat, cow,zebrafish
  - mRNAs
  - predicted mRNAs



UniGene

UniGene Human Sequences Collection

Search using text

At the top of most UniGene pages you will see a text field for use in formulating queries. Query terms may be ordinary words like "kinase" or specific identifiers, such as GenBank accession numbers. In addition, a number of @functions are provided for specialized purposes.

For more information, see our Query Tips.

Search by Chromosome

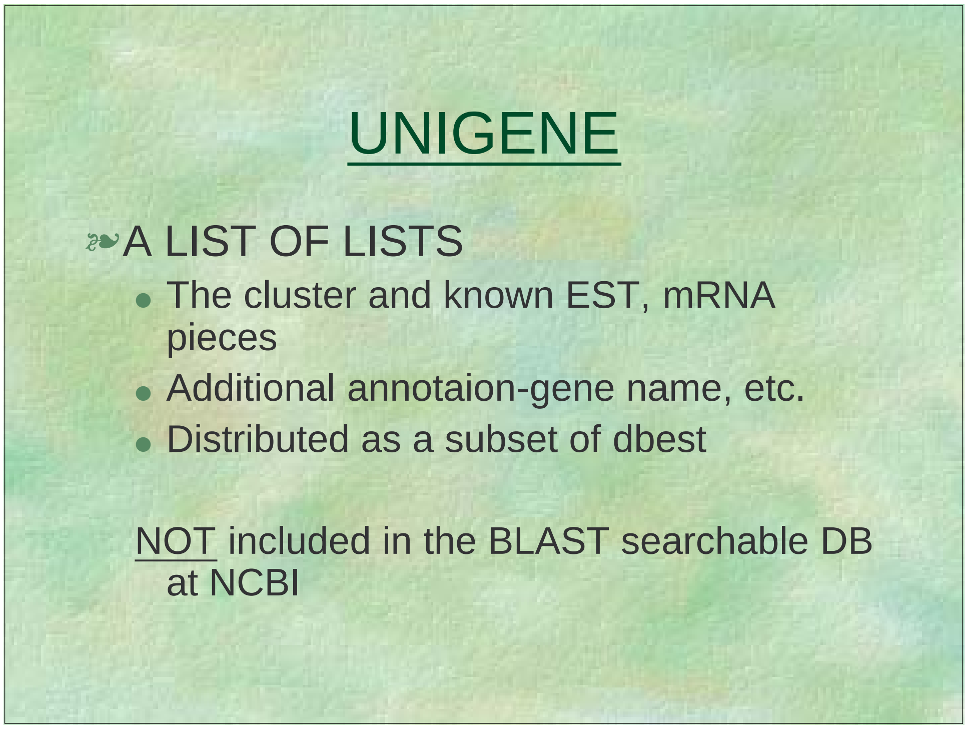
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y Unk

Information on map position has been imported from several sources, including OMIM, the RH Consortium Human Transcript Map, and the Whitehead Institute human physical map. Click on a chromosome number (above) to see a list of UniGene entries for that chromosome. Alternatively, you may use **@chr(num)** as a search term.

Search by Library

Use the Library Browser to see a list of cDNA libraries that have been used in EST projects. Note that each library has an ID number, which may also be used within a text query as **@lib(num)**.

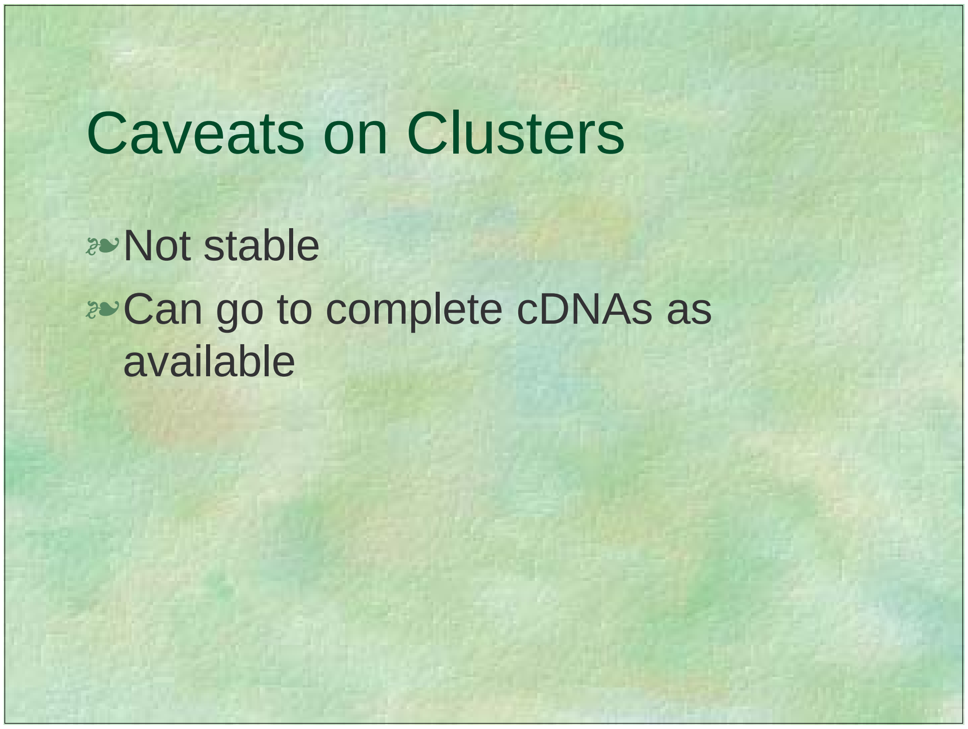
NLM | NIH | UniGene | Privacy Statement | Disclaimer | NCBI Help



UNIGENE

- A LIST OF LISTS
  - The cluster and known EST, mRNA pieces
  - Additional annotaion-gene name, etc.
  - Distributed as a subset of dbest

NOT included in the BLAST searchable DB at NCBI



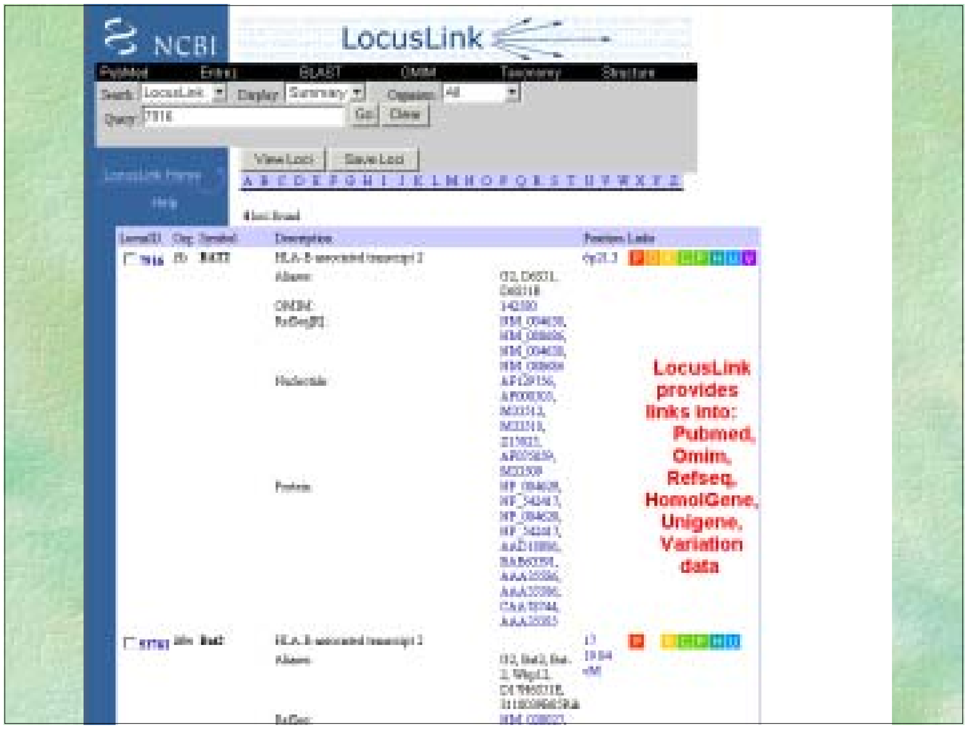
Caveats on Clusters

- Not stable
- Can go to complete cDNAs as available

LOCUSLINK

(http://www.ncbi.nlm.nih.gov/LocusLink)

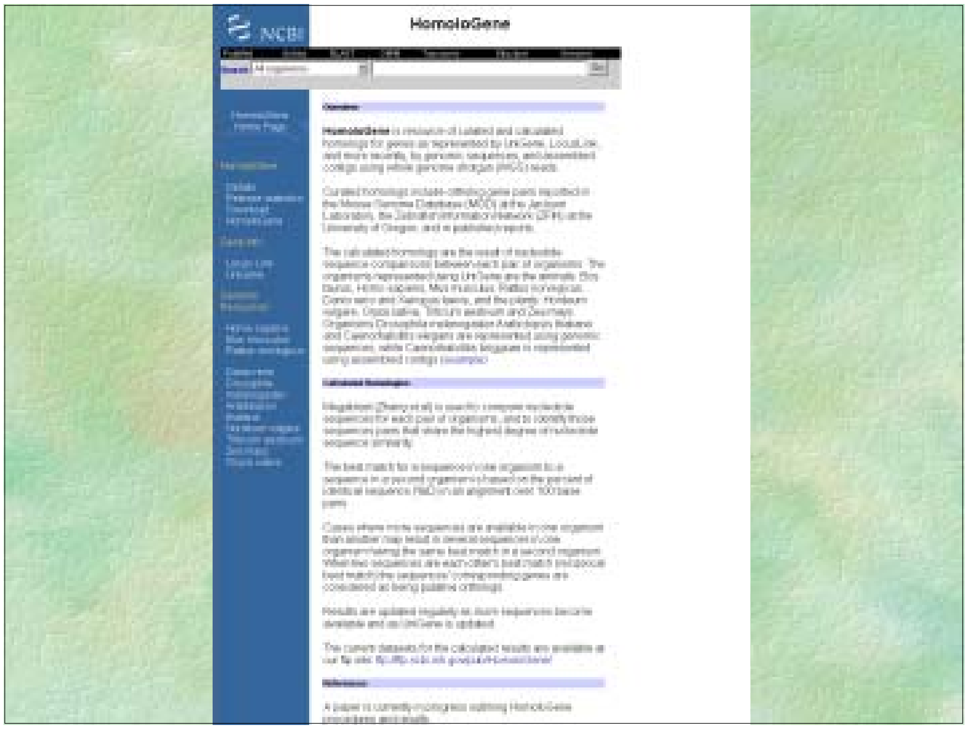
- A useful, searchable compendium of loci across human, mouse, rat, Drosophila and zebrafish
- Linked for PubMed, OMIM, RefSeq, Homologene data, Unigene, and Variation Data


NCBI
LocusLink
PubMed Entrez BLAST OMIM Taxonomy Structure
Search: LocusLink Display: Summary Organism: All
Query: 7916 Go Clear
View Loci Save Loci
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
LocusLink Home
Help

4 loci found

LocusID	Org	Symbol	Description		Position	Links
□ 7916	Hs	BAT2	HLA-B associated transcript 2		6p21.3	P O R G H U V
			Aliases:	G2, D6S51, D6S51E		
			OMIM:	142580		
			RefSeq[R]:	NM_004638, NM_080686, NM_004638, NM_080686		
			Nucleotide:	AF129756, AP000505, M33512, M33518, Z15025, AF075839, M33509		
			Protein:	NP_004629, NP_542417, NP_004629, NP_542417, AAD18086, BAB63391, AAA35586, AAA35590, CAA78744, AAA35583		
□ 53761	Mm	Bat2	HLA-B associated transcript 2		17 19.04 cM	P G H U
			Aliases:	G2, Bat2, Bat-2, Wbp12, D17H6S51E, 3110039K05Rik		
			RefSeq:	NM_020027,		

LocusLink provides links into: Pubmed, Omim, Refseq, HomolGene, Unigene, Variation data



Overview

HomoloGene is a resource of curated and calculated homologs for genes as represented by UniGene, LocusLink, and more recently, by genomic sequences, and assembled contigs using whole genome shotgun (WGS) reads.

Curated homologs include orthologous gene pairs reported in the Mouse Genome Database (MGD) at the Jackson Laboratory, the Zebrafish Information Network (ZFIN) at the University of Oregon, and in published reports.

The calculated homologs are the result of nucleotide sequence comparisons between each pair of organisms. The organisms represented using UniGene are the animals: Bos taurus, Homo sapiens, Mus musculus, Rattus norvegicus, Danio rerio and Xenopus laevis, and the plants: Hordeum vulgare, Oryza sativa, Triticum aestivum and Zea mays. Organisms Drosophila melanogaster, Arabidopsis thaliana and Caenorhabditis elegans are represented using genomic sequences, while Caenorhabditis briggsae is represented using assembled contigs (example).

Calculated Homologs

MegaBlast (Zhang et al) is used to compare nucleotide sequences for each pair of organisms, and to identify those sequence pairs that share the highest degree of nucleotide sequence similarity.

The best match for a sequence in one organism to a sequence in a second organism is based on the percent of identical sequence (%ID) in an alignment over 100 base pairs.

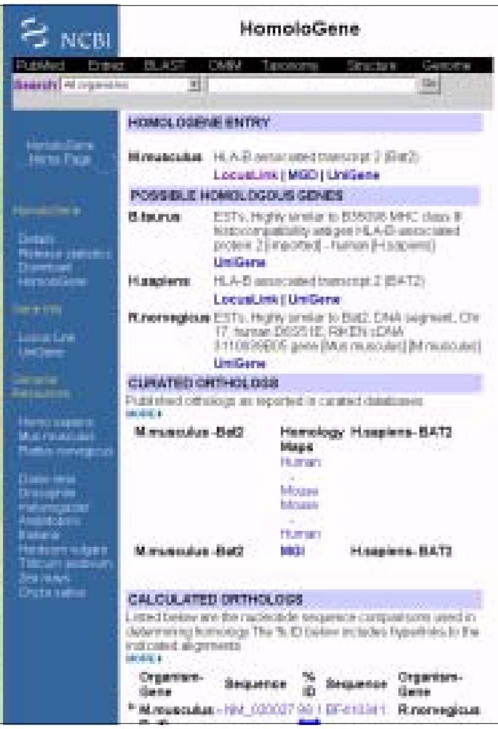
Cases where more sequences are available in one organism than another may result in several sequences in one organism having the same best match in a second organism. When two sequences are each other's best match (reciprocal best match) the sequences' corresponding genes are considered as being putative orthologs.

Results are updated regularly as more sequences become available and as UniGene is updated.

The current datasets for the calculated results are available at our ftp site: ftp://ftp.ncbi.nih.gov/pub/HomoloGene/

Reference

A paper is currently in progress outlining HomoloGene procedures and results.



HOMOLOGENE ENTRY

M.musculus HLA-B associated transcript 2 (Bat2)
LocusLink | MGD | UniGene

POSSIBLE HOMOLOGOUS GENES

B.taurus ESTs, Highly similar to B35098 MHC class III histocompatibility antigen HLA-B-associated protein 2 [imported] - human [H.sapiens]
UniGene

H.sapiens HLA-B associated transcript 2 (BAT2)
LocusLink | UniGene

R.norvegicus ESTs, Highly similar to Bat2; DNA segment, Chr 17, human D6S51E; RIKEN cDNA 3110039B05 gene [Mus musculus] [M.musculus]
UniGene

CURATED ORTHOLOGS

Published orthologs as reported in curated databases
details

M.musculus -Bat2	Homology Maps Human - Mouse Mouse - Human	H.sapiens- BAT2
M.musculus -Bat2	MGI	H.sapiens- BAT2

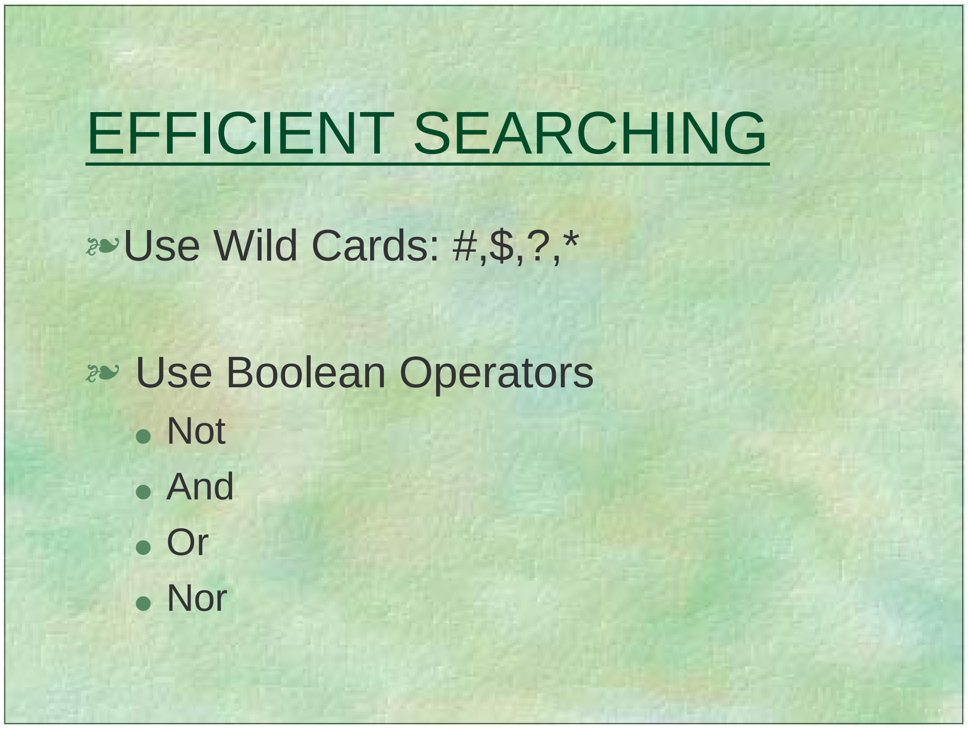
CALCULATED ORTHOLOGS

Listed below are the nucleotide sequence comparisons used in determining homology. The % ID below includes hyperlinks to the indicated alignments.
details

	Organism-Gene	Sequence	% ID	Sequence	Organism-Gene
*	M.musculus -Bat2	NM_020027	90.1	BF410341	R.norvegicus

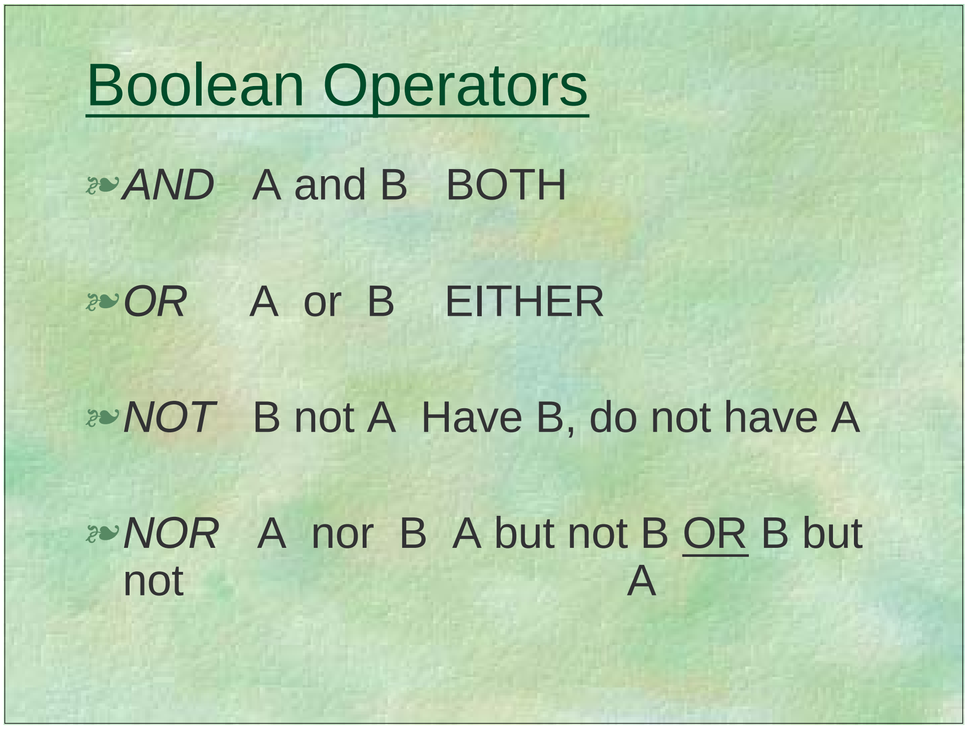
Resources for Genomic Comparison

- GLASS-http://plover.lcs.mit.edu
- PipMaker: http://bio.cse.psu.edu
- Rosetta: http:// plover.lcs.mit.edu(genes)
- SGP: htttp://soft.ice.mpg.de/sgp-1
- VISTA: http://www-gsd.lbl.gov/VISTA
- WABA: http://www.cse.ucsc.edu/~kent/xenoAli/index.html



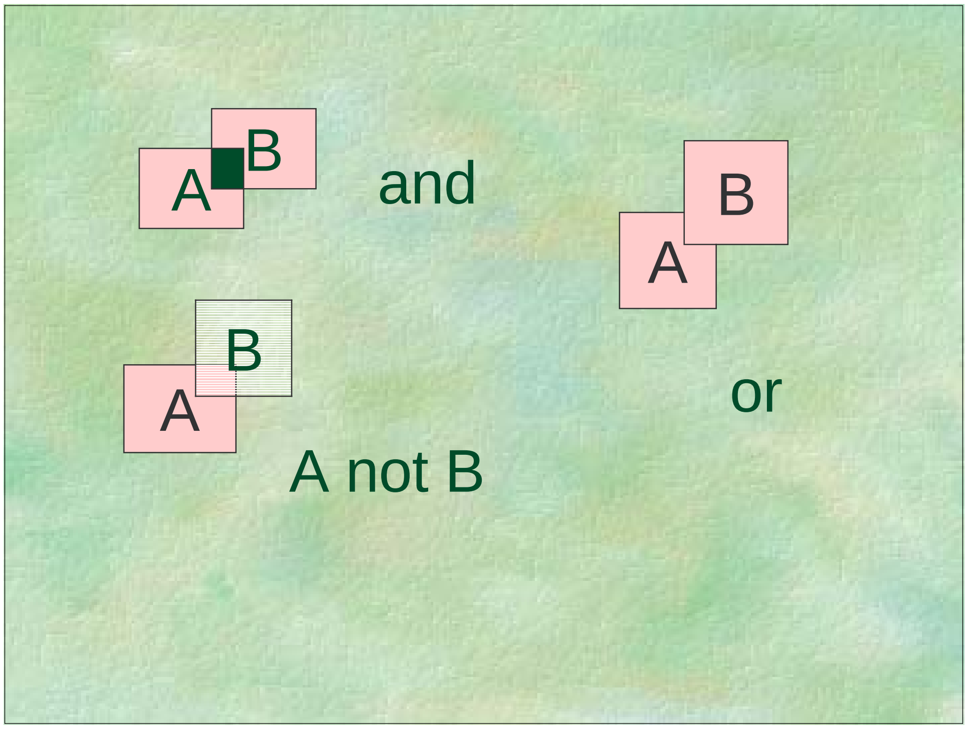
EFFICIENT SEARCHING

- Use Wild Cards: #,$,?,*
- Use Boolean Operators
  - Not
  - And
  - Or
  - Nor



Boolean Operators

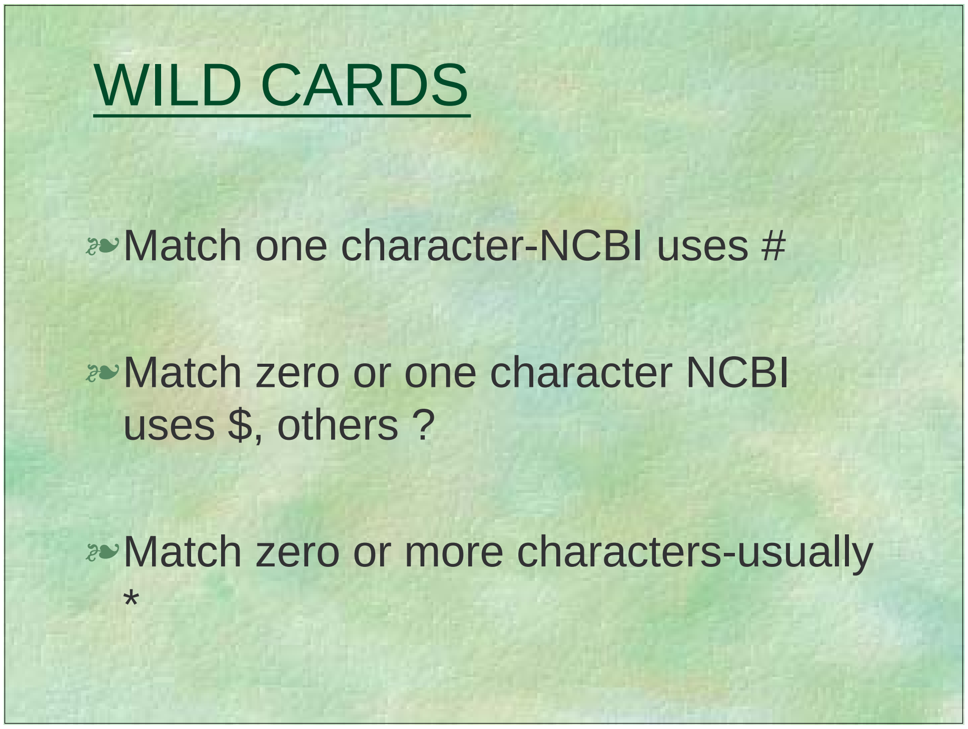
- *AND* A and B BOTH
- *OR* A or B EITHER
- *NOT* B not A Have B, do not have A
- *NOR* A nor B A but not B OR B but
  not A



A B and

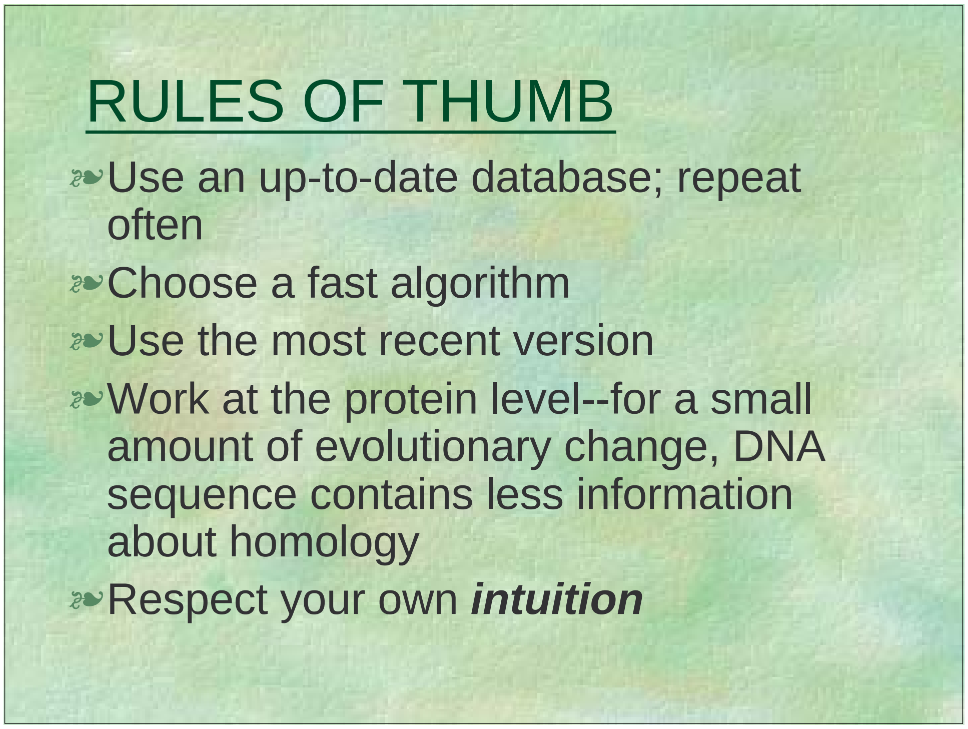
A B or

A B A not B



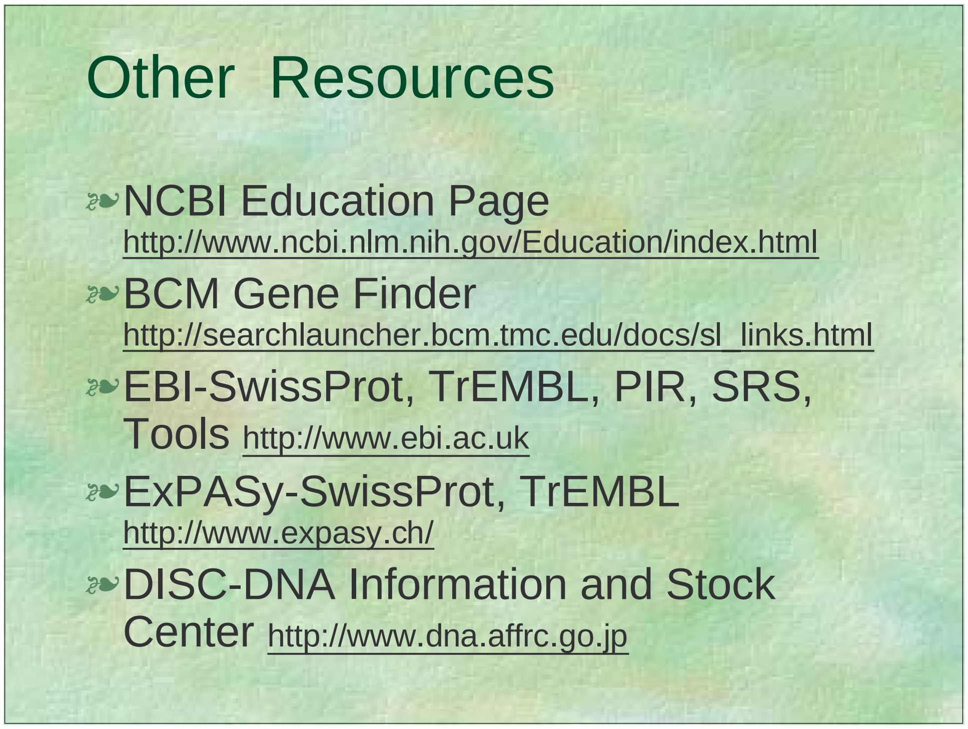
WILD CARDS

- Match one character-NCBI uses #
- Match zero or one character NCBI uses $, others ?
- Match zero or more characters-usually *



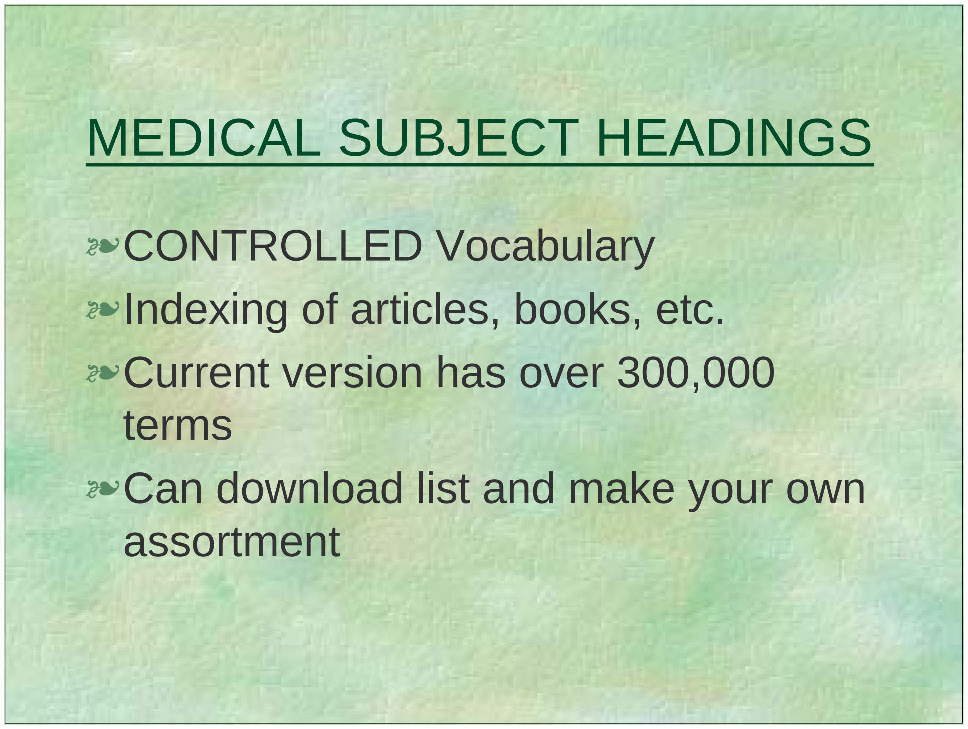
RULES OF THUMB

- Use an up-to-date database; repeat often
- Choose a fast algorithm
- Use the most recent version
- Work at the protein level--for a small amount of evolutionary change, DNA sequence contains less information about homology
- Respect your own ***intuition***



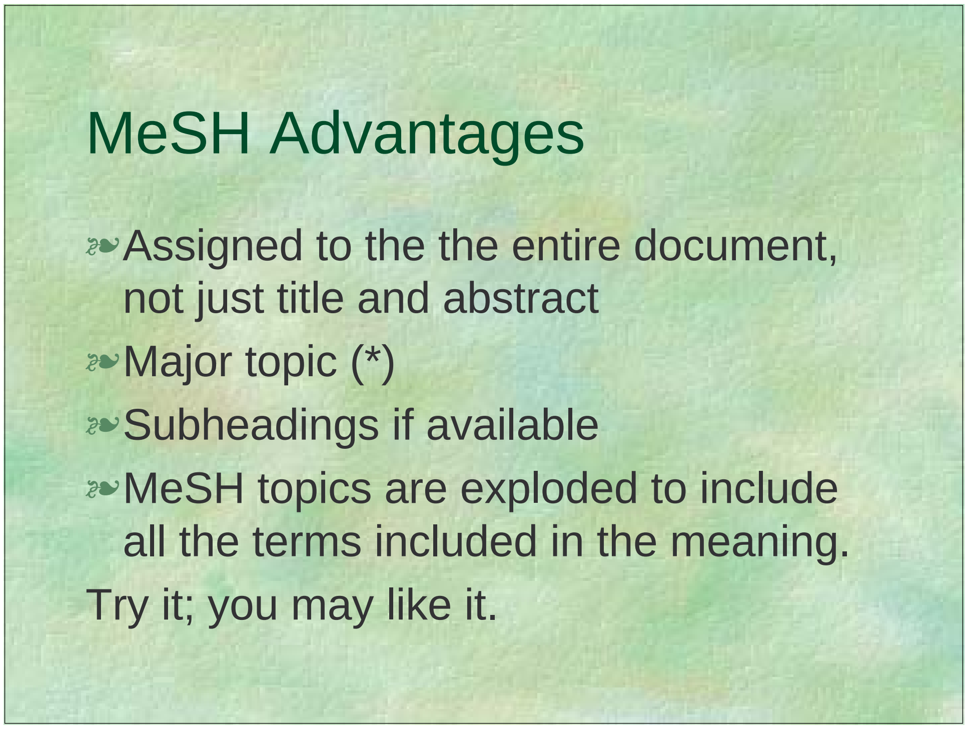
Other Resources

- NCBI Education Page http://www.ncbi.nlm.nih.gov/Education/index.html
- BCM Gene Finder http://searchlauncher.bcm.tmc.edu/docs/sl_links.html
- EBI-SwissProt, TrEMBL, PIR, SRS, Tools http://www.ebi.ac.uk
- ExPASy-SwissProt, TrEMBL http://www.expasy.ch/
- DISC-DNA Information and Stock Center http://www.dna.affrc.go.jp



MEDICAL SUBJECT HEADINGS

- CONTROLLED Vocabulary
- Indexing of articles, books, etc.
- Current version has over 300,000 terms
- Can download list and make your own assortment

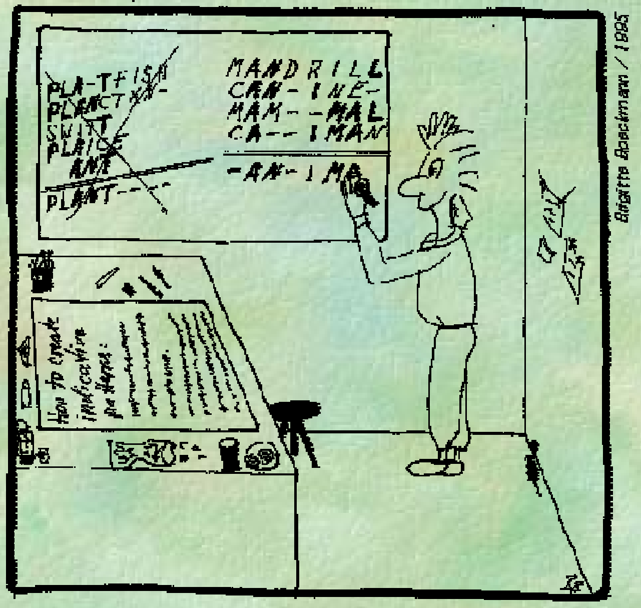


MeSH Advantages

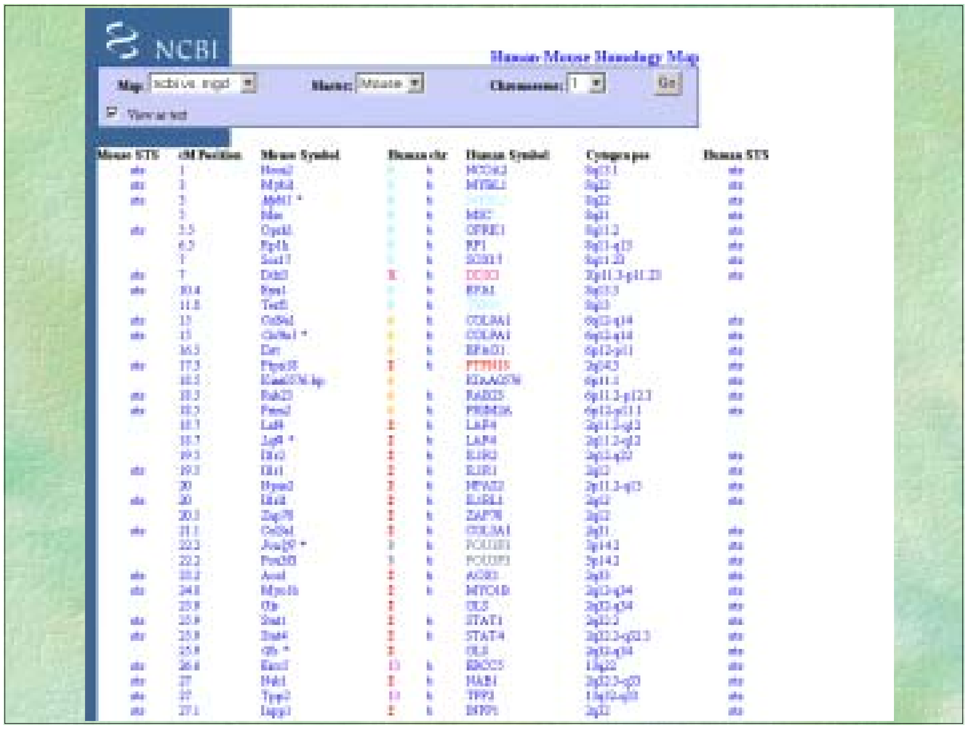
- Assigned to the the entire document, not just title and abstract
- Major topic (*)
- Subheadings if available
- MeSH topics are exploded to include all the terms included in the meaning.

Try it; you may like it.

How we develop Prosite patterns!



"'Potential', 'Probable', 'By Similarity' -
They don't know anything about this protein!"



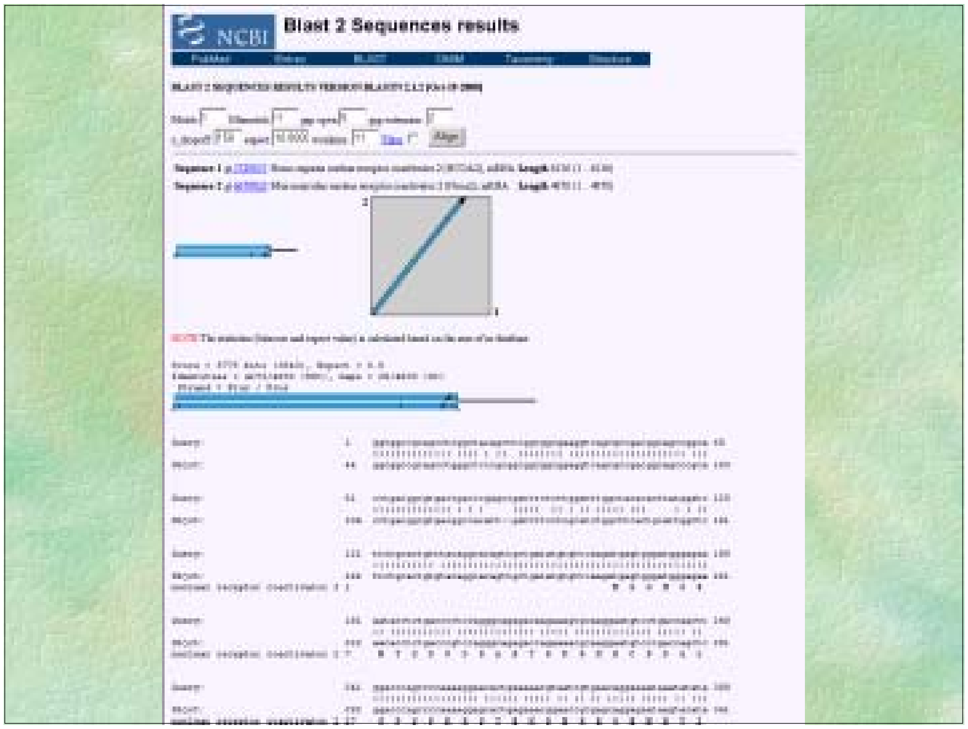
NCBI

Human-Mouse Homology Map

Map: ncbi vs mgd ▾ Master: Mouse ▾ Chromosome: 1 ▾ Go

☑ View as text

Mouse STS	cM Position	Mouse Symbol	Human chr		Human Symbol	Cytogen pos	Human STS
sts	1	Ncoa2	8	h	NCOA2	8q13.1	sts
sts	1	Mybl1	8	h	MYBL1	8q22	sts
sts	3	[illegible] *	8	h	[illegible]	8q22	sts
	3	Mos	8	h	MOS	8q11	sts
sts	3.5	Oprk1	8	h	OPRK1	8q11.2	sts
	6.3	Rp1h	8	h	RP1	8q11-q13	sts
	7	Sox17	8	h	SOX17	8q11.23	sts
sts	7	Ddx3	X	h	DDX3	Xp11.3-p11.23	sts
sts	10.4	Rrs1	8	h	RPA1	8q13.3	
	11.8	Tcea1	8	h	[illegible]	8q11	
sts	13	Col9a1	6	h	COL9A1	6q12-q14	sts
sts	13	Col9a1 *	6	h	COL9A1	6q12-q14	sts
	16.5	Dst	6	h	BPAG1	6p12-p11	sts
sts	17.3	Ptpn18	2	h	PTPN18	2q14.3	sts
	18.3	Kiaa0576-hp	6		KIAA0576	6p11.1	sts
sts	18.3	Rab23	6	h	RAB23	6p11.2-p12.3	sts
sts	18.3	Prim2	6	h	PRIM2A	6p12-p11.1	sts
	18.7	Laf4	2	h	LAF4	2q11.2-q12	
	18.7	[illegible] *	2	h	LAF4	2q11.2-q12	
	19.3	Il1r2	2	h	IL1R2	2q12-q22	sts
sts	19.3	Il1r1	2	h	IL1R1	2q12	sts
	20	Hpsd2	2	h	MPAD2	2p11.2-q13	sts
sts	20	Il1rl1	2	h	IL1RL1	2q12	sts
	20.3	Zap70	2	h	ZAP70	2q12	
sts	21.1	Col3a1	2	h	COL3A1	2q31	sts
	22.2	[illegible] *	3	h	POU3F3	3p14.2	sts
	22.2	Pou3f3	3	h	POU3F3	3p14.2	sts
sts	23.2	Aox1	2	h	AOX1	2q33	sts
sts	24.8	Myo1b	2	h	MYO1B	2q12-q34	sts
	25.8	Gls	2		GLS	2q32-q34	sts
sts	25.8	Stat1	2	h	STAT1	2q32.2	sts
sts	25.8	Stat4	2	h	STAT4	2q32.2-q32.3	sts
	25.8	Gls *	2		GLS	2q32-q34	sts
sts	26.8	Ercc5	13	h	ERCC5	13q22	sts
sts	27	Nab1	2	h	NAB1	2q32.3-q33	sts
sts	27	Tpp2	13	h	TPP2	13q32-q33	sts
sts	27.1	Inpp1	2	h	INPP1	2q32	sts



NCBI

Blast 2 Sequences results

BLAST 2 SEQUENCES RESULTS VERSION BLASTN 2.1.2 [Oct-19-2000]

Match [ ] Mismatch [ ] gap open [ ] gap extension [ ]

x_dropoff [ ] expect [ ] wordsize [ ] Filter □ Align

Sequence 1 gi [illegible] Homo sapiens nuclear receptor coactivator 2 (NCOA2), mRNA Length [illegible]

Sequence 2 gi [illegible] Mus musculus nuclear receptor coactivator 2 (Ncoa2), mRNA Length [illegible]

NOTE: The statistics (bitscore and expect value) is calculated based on the size of nr database

```
Score = [illegible]
Identities = [illegible], Gaps = [illegible]
Strand = Plus / Plus

Query:                          1   [illegible]  60
Sbjct:                         44   [illegible]  103

Query:                         61   [illegible]  120
Sbjct:                        104   [illegible]  163

Query:                        121   [illegible]  180
Sbjct:                        164   [illegible]  223
nuclear receptor coactivator 2 1

Query:                        181   [illegible]  240
Sbjct:                        224   [illegible]  283
nuclear receptor coactivator 2 7

Query:                        241   [illegible]  300
Sbjct:                        284   [illegible]  343
nuclear receptor coactivator 2 27
```